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

Will E. Thompson, MS[1], **David M. Vidmar, PhD**[1], **Jessica K. De Freitas, PhD**[1],
John M. Pfeifer, MD, MPH[1], **Brandon K. Fornwalt, MD, PhD**[1], **Ruijun Chen, MD, MS**[1],
Gabriel Altay, PhD[1], **Kabir Manghnani, BS**[1], **Andrew C. Nelsen, PharmD**[2],
Kellie Morland, PharmD[2], **Martin C. Stumpe, PhD**[1], **Riccardo Miotto, PhD**[1]

[1]Tempus Labs, Inc., 600 West Chicago Avenue, Suite 510, Chicago, IL 60654 USA
[2]United Therapeutics Corporation, 1000 Spring Street, Silver Spring, Maryland 20910 USA

Abstract

Identifying disease phenotypes from electronic health records (EHRs) is critical for numerous secondary uses. Manually encoding physician knowledge into rules is particularly challenging for rare diseases due to inadequate EHR coding, necessitating review of clinical notes. Large language models (LLMs) offer promise in text understanding but may not efficiently handle real-world clinical documentation. We propose a zero-shot LLM-based method enriched by retrieval-augmented generation and MapReduce, which pre-identifies disease-related text snippets to be used in parallel as queries for the LLM to establish diagnosis. We show that this method as applied to pulmonary hypertension (PH), a rare disease characterized by elevated arterial pressures in the lungs, significantly outperforms physician logic rules (F_1 score of 0.62 vs. 0.75). This method has the potential to enhance rare disease cohort identification, expanding the scope of robust clinical research and care gap identification.

1 Introduction

Disease phenotyping within electronic health records (EHRs) involves identifying ground truth diagnoses in a patient's clinical history. These phenotypes play a crucial role in several essential functions, such as selecting patient groups for observational studies or interventional quality initiatives to close gaps in care, defining inclusion and exclusion criteria, and providing labels for subsequent modeling tasks. Relevant information for disease diagnosis may be scattered across EHRs different data sources, including free-text notes, International Classification of Diseases 9th and 10th revision (ICD-9/10 CM) codes, medications, or laboratory values from medical procedures and tests.

The ideal process involves subject matter experts (SMEs) reviewing patient files for disease diagnosis. Yet chart reviews are time-consuming, taking an average of 30 minutes per file [1]. To address this, SMEs often create rules-based algorithms, combining ICD codes, laboratory values, medications, and procedures, to identify diseases. However, challenges arise, including coding errors, reporting biases, and data sparsity, requiring iterative refinement through a human-in-the-loop process. Scalability is hindered, especially when features from one EHR system do not generalize to others [2]. Mapping rare diseases to common ontologies can be especially problematic due to expert disagreements [3, 4].

Machine learning phenotyping approaches, both supervised and unsupervised, have shown varying degrees of promise [5, 6, 7, 8, 9, 10]. Supervised learning approaches often require high-quality labels and are therefore constrained by a labeling bottleneck. While unsupervised learning approaches circumvent this problem, they often are difficult to tailor to a particular disease definition or achieve certain acceptance criteria. The majority of work also mostly focuses on structured EHRs; incor-

porating the notes in a phenotype definition holds the possibility of developing a better model [11, 12].

To overcome these limitations within practical computational constraints, this paper introduces a retrieval-augmented generation (RAG) approach to zero-shot phenotyping using large language models (LLMs). To our knowledge, this is the first study to explore the application of a RAG approach to process entire patient records with LLMs, as opposed to focusing solely on a specific type of clinical note. An essential aspect of this retrieval step is its ability to analyze all clinical mentions throughout a patient's entire record without the need for predefined sections of interest. Given the substantial volume of potentially relevant information retrieved, and the aim to assess the language model's reasoning abilities rather than relying on an intricate retrieval mechanism, this paper also investigates a MapReduce paradigm for parallel snippet evaluation and resolution of potentially conflicting information during the output aggregation stage. To assess the performance of this approach, we enlist a physician SME to assist in developing a competing rules-based model, the commonly used approach in healthcare practice and industry. Both models are then evaluated using an unseen test dataset, with our chosen disease phenotype being pulmonary hypertension (PH).

2   Methodology

LLM Phenotyping Pipeline   Figure 1 shows the overall architecture of the LLM-based phenotyping model: a RAG approach followed by a MapReduce step for querying and aggregation.
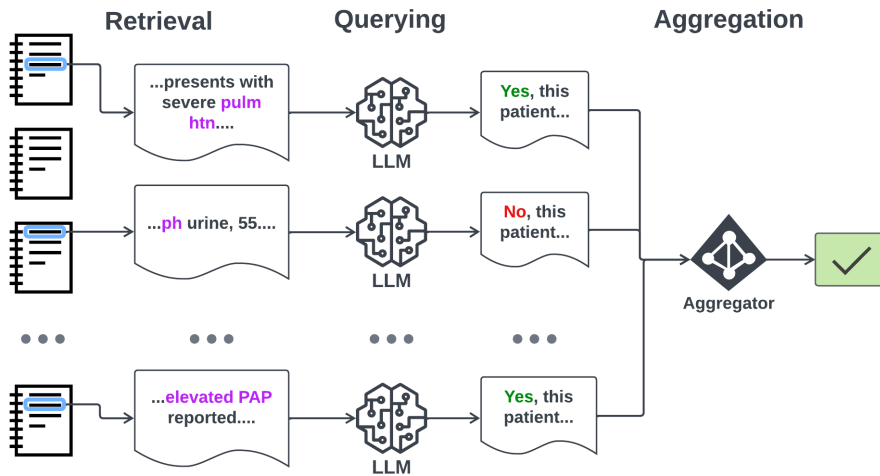

Figure 1: Overview of the LLM-based phenotype architecture.

Retrieval   Considering the extensive volume of text contained within a real world data (RWD) warehouse of EHRs, it becomes impractical to process the entirety of a patient's clinical notes within the context window of an LLM. To address this challenge, we implemented a RAG approach based on regular expressions (Regex) to identify relevant snippets from the text. As shown in the literature, a RAG approach proves to be more efficient and effective than providing the LLM with larger context windows [13]. We opted for Regex over more complex retrieval models, such as Term Frequency-Inverse Document Frequency, Cohere's re-rank [14], or Instructor embeddings [15], in order to avoid introducing bias through additional hyperparameter tuning and to narrow the focus of our work to assessing the LLM's capability in diagnosing diseases. Although building Regex queries requires domain knowledge which may not be easily extensible to all other disease contexts, we believe that this approach injects useful priors into the model pipeline. Further, some form of retrieval method is necessary as passing the entirety of a patients' clinical notes through an LLM is *intractable* given the sheer size of clinical notes, even in the context of a MapReduce paradigm. We divided each clinical note into snippets of 2,048 "tokens" (i.e., the vocabulary of characters, words or subwords an LLM uses to partition free text) in size and then employed Regex to identify relevant snippets. Regex rules encompassed a broad spectrum of patterns that could potentially be associated

with any mention of a specific medical condition, such as PH (Appendix A.1). Subsequently, all Regex-retrieved patient snippets are fed into the LLM to determine the diagnosis status.

Querying   While the RAG approach reduces the amount of text processed by the LLM, RWD clinical notes often comprise many pages of text. Consequently, the Regex retriever is still likely to return a large number of snippets, which may exceed the LLM's context window. To address this, we employed a MapReduce approach where each snippet is concurrently presented as context to the LLM, along with a set of instructions to facilitate decision-making (i.e., the "map" phase). Generated outputs, one per snippet, are then aggregated to form the final decision (i.e., the "reduce" phase).

Aggregation   We evaluated two aggregation approaches for the "reduce" phase: (1) an LLM-based approach, which aggregates the output and reasoning from each individual snippet query into a larger prompt for a final decision by an LLM through prompting; and (2) a Max aggregation function, which checks if any of the individual snippet queries returned a positive diagnosis and, if so, assigns a positive label to the patient as a whole. In the LLM aggregation approach, we explored two variations: (1) apply the same prompt that was provided at the snippet level to aggregate responses; and (2) apply a different prompt that asks the LLM if any of the responses indicated a positive diagnosis.

3   Evaluation Design

Pulmonary Hypertension   PH is a condition characterized by abnormally elevated pressure in the arteries of the lung and right side of the heart. While the prevalence of specific PH etiologies may vary across groups, it is broadly categorized as a rare disease with an estimated global prevalence rate of 1-3% [16]. Building a PH phenotype is complicated by the fact that the hemodynamic definition of PH has changed over time [17]. This means that there are some patients who would be currently identified as having PH under the new definition, would not have been diagnosed with PH during the time of their original treatment or workup. The ability to systematically identify PH patients who would otherwise not be identified could significantly impact patient outcomes.

Dataset   We utilized de-identified clinical notes from a medium-sized health system serving a population of approximately 2.2 million patients (as of December 2022). Given the expected low prevalence rate of PH within this population, we created an enriched cohort of patients displaying any clinical evidence of PH in either the structured data or clinical notes. From this cohort, we randomly selected 299 patients for a comprehensive chart review[1]. Each patient underwent independent evaluation by two physicians, with any discrepancies resolved through joint discussion to reach a consensus. The cohort consisted primarily of a white population (97%), comprised of 154 females and 145 males. The median age at the last encounter for each patient was 72 years (interquartile range: 60 to 84), and the median observation time, defined as the number of years between each patient's first and last encounter, was 17.5 (interquartile range: 8.75 to 22.92). Subsequently, these labeled patients were divided into three groups: a training set comprising 50 patients (19 cases and 31 controls); a validation set also consisting of 50 patients (19 cases and 31 controls); and a test set with 199 patients (76 cases and 123 controls).

Large Language Model   We utilized 'bison@001', which is a version of PaLM-2 available through Google Cloud's Vertex-AI offering [18]. As of the time of writing, this stands as the second-largest commercially available variant of the PaLM-2 model intended for consumer use. PaLM-2 builds upon the foundation of PaLM-1, incorporating a combination of various pre-training objectives to achieve state-of-the-art results on several benchmarks while maintaining smaller model sizes [19].

Structured Phenotype Definition   One of our internal physicians conducted a review of patients within the training dataset to establish a rules-based algorithm for diagnosing PH using EHRs. Following a thorough examination of the literature on PH phenotypes, the rules encompassed a blend of ICD-9/10 code frequencies, medication records, laboratory data, and other clinical features available in the patients' records. After a series of iterative reviews and adjustments to the model output, the physician ceased further model development when the incremental improvements began to diminish. The final phenotype is reported in Appendix A.2 Table 2.

---

[1]We originally sampled 300 patients for curation; however, due to data integration issues, we dropped one patient during the curation process.

4   Results

We conducted experiments involving several iterations of prompt design, snippet exclusions and aggregation methods. After evaluating the results using training and validation sets, we selected a subset of configurations for final assessment on the test set. The complete set of results on the validation set is presented in Appendix A.3 Table 3. All performances are in terms of F_1 score.

Prompt Design Analysis   We explored various zero-shot prompt designs to query the LLM for the diagnosis of PH. Additionally, we evaluated the value of Chain-of-Thought (CoT) reasoning by enhancing the prompt with the phrase "let's think step-by-step" [20, 21]. False positives and false negatives were manually reviewed while iterating on prompt design. From this analysis, we concluded that our prompts needed to explicitly guide the model to consider possible cases of PH as negative diagnosis and history of PH as positive diagnosis. We found a fair amount of variability in performance across these different prompt designs without any prominent features defining those prompts that appeared to perform best. Therefore we chose the prompt design presented in Figure 2, which is composed of CoT, multiple choice, history of PH and possible PH steering. See Appendix A.3 Table 4 for the description of each prompt evaluated.


```
### 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.
###
```

Chain of Thought (CoT)
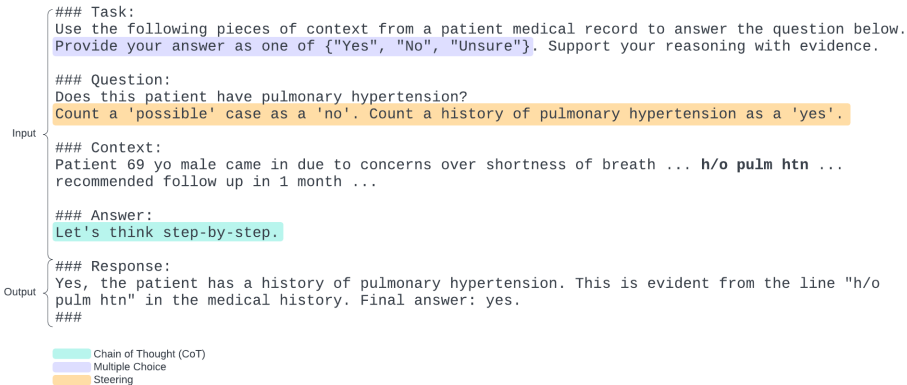Multiple Choice
Steering

Figure 2: An illustrative example of zero-shot prompt.

Retrieval Exclusions   While reviewing model errors on the validation set, we observed that a noteworthy percentage of false positives originated from echocardiogram (ECHO) or computerized tomography (CT) reports noting *suspicion* of PH without any confirmatory diagnostic testing nor a clinical diagnosis by a provider. This was not surprising since it is known that both ECHO and CT have higher error rates when compared to the gold standard diagnostic i.e. right heart catheterization (RHC) [22, 23, 24, 25]. As these reports typically exhibit a consistent structure, we explored the possibility of excluding snippets extracted from the reports in two ways: (1) employing regular expressions to filter out snippets containing headers and common language found in these reports and (2) enhancing our prompts to instruct the LLM to disregard ECHO and CT reports. Results on the validation set indicated improvements in the performance of the model when excluding these snippets; therefore, we decided to include this filter in the final architecture. We also noticed that the structured phenotype definition did not include ECHO reports as it was built solely using structured EHRs and not clinical notes.

Aggregation Evaluation   LLM-based aggregation methods, whether using the same or different prompts at the snippet level, resulted in average F_1 scores of 0.67 and 0.68, respectively, with significant variability in performance across choice of prompting design. In contrast, Max aggregation achieved an F_1 score of 0.73, also providing increased stability across different prompts.

Comparing LLM and Structured Phenotypes   Table 1 compares the performance of three variations of the LLM-based phenotyping architecture with the structured phenotype baseline developed by a physician on the test set. As demonstrated, LLM-based phenotypes generally show improvements between 18% and 21% over the structured phenotype. We did observe a notable drop in F_1

scores, ranging from 0.05 to 0.1, in the performance of LLM-based methods compared to the results obtained on the validation set (see Appendix A.2 Table 2), which might be attributed to the larger evaluation cohort and potentially to some overfitting on the training set. Nevertheless, the LLM-based methods significantly outperformed the structured phenotype method, resulting in the identification of approximately twice as many patients with a confirmed diagnosis of PH. In a real-world application, these patients might otherwise remain unidentified. Furthermore, we observed that our retrieved documents spanned 29 distinct note types, highlighting the importance of retrieving across note types to accurately identify disease diagnoses (see Appendix A.4 Table 5).

Model	Aggregation	ECHO Exclusion	F_1 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 of three variants of the LLM-based phenotype compared to structured phenotype.

5   Conclusion

This paper underscores the potential of employing an LLM-based architecture to identify diseases across clinical notes. Unlike existing literature, which often utilizes LLMs on specific types of notes, our method harnesses RAG and MapReduce to effectively analyze the complete patient documentation. Our experiments demonstrated the superiority of this method over SME rule-based models in diagnosing PH. Efficient LLM-based phenotype models offer scalability and improvement in identifying specific diseases in real-world EHRs, reducing the manual workload for SMEs and the need for ad-hoc machine learning models while enabling comprehensive patient record analysis. This advancement promises to enhance systems utilizing EHRs for purposes such as clinical decision support, care gap detection / population health management, clinical trial matching, and cohort generation. Future work will explore the integration of this method with structured data, the evaluation of other LLM engines, SME-enabled active learning loops, more advanced retrieval methods and additional clinical domains.

References

[1] Jordan McKenzie et al. "A Semiautomated Chart Review for Assessing the Development of Radiation Pneumonitis Using Natural Language Processing: Diagnostic Accuracy and Feasibility Study". In: *JMIR Medical Informatics* 9.11 (Nov. 2021), e29241. ISSN: 2291-9694. DOI: 10.2196/29241. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8663661/ (visited on 09/11/2023).

[2] Anna Ostropolets et al. "Reproducible variability: assessing investigator discordance across 9 research teams attempting to reproduce the same observational study". eng. In: *Journal of the American Medical Informatics Association: JAMIA* 30.5 (Apr. 2023), pp. 859–868. ISSN: 1527-974X. DOI: 10.1093/jamia/ocad009.

[3] Kin Wah Fung, Rachel Richesson, and Olivier Bodenreider. "Coverage of Rare Disease Names in Standard Terminologies and Implications for Patients, Providers, and Research". In: *AMIA Annual Symposium Proceedings* 2014 (Nov. 2014), pp. 564–572. ISSN: 1942-597X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419993/ (visited on 09/11/2023).

[4] Jonathan M. Mortensen, Mark A. Musen, and Natalya F. Noy. "An empirically derived taxonomy of errors in SNOMED CT". In: *AMIA Annual Symposium Proceedings* 2014 (Nov. 2014), pp. 899–906. ISSN: 1942-597X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419962/ (visited on 09/11/2023).

[5] Joyce C. Ho et al. "Limestone: high-throughput candidate phenotype generation via tensor factorization". eng. In: *Journal of Biomedical Informatics* 52 (Dec. 2014), pp. 199–211. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2014.07.001.

[6] Rimma Pivovarov et al. "Learning probabilistic phenotypes from heterogeneous EHR data". eng. In: *Journal of Biomedical Informatics* 58 (Dec. 2015), pp. 156–165. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2015.10.001.

[7] Yoni Halpern et al. "Electronic medical record phenotyping using the anchor and learn framework". eng. In: *Journal of the American Medical Informatics Association: JAMIA* 23.4 (July 2016), pp. 731–740. ISSN: 1527-974X. DOI: 10.1093/jamia/ocw011.

[8] Benjamin S. Glicksberg et al. "Automated disease cohort selection using word embeddings from Electronic Health Records". eng. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 23 (2018), pp. 145–156. ISSN: 2335-6936.

[9] Yuri Ahuja et al. "sureLDA: A multidisease automated phenotyping method for the electronic health record". In: *Journal of the American Medical Informatics Association* 27.8 (Aug. 2020), pp. 1235–1243. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa079. URL: https://doi.org/10.1093/jamia/ocaa079 (visited on 09/11/2023).

[10] Jessica K. De Freitas et al. "Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records". In: *Patterns* 2.9 (Sept. 2021), p. 100337. ISSN: 2666-3899. DOI: 10.1016/j.patter.2021.100337. URL: https://www.sciencedirect.com/science/article/pii/S2666389921001859 (visited on 09/11/2023).

[11] David Vidmar et al. "Abstract 11934: Natural Language Processing Models Can Be Trained to Accurately Recognize the Presence of Disease Within Clinical Notes". In: *Circulation* 146.Suppl_1 (Nov. 2022). Publisher: American Heart Association, A11934–A11934. DOI: 10.1161/circ.146.suppl_1.11934. URL: https://www.ahajournals.org/doi/10.1161/circ.146.suppl_1.11934 (visited on 09/21/2023).

[12] Asher Moldwin, Dina Demner-Fushman, and Travis R. Goodwin. "Empirical Findings on the Role of Structured Data, Unstructured Data, and their Combination for Automatic Clinical Phenotyping". In: *AMIA Summits on Translational Science Proceedings* 2021 (May 2021), pp. 445–454. ISSN: 2153-4063. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8378600/ (visited on 09/21/2023).

[13] Amnon Catav. *Less is More: Why Use Retrieval Instead of Larger Context Windows | Pinecone*. en. July 2023. URL: https://www.pinecone.io/blog/why-use-retrieval-instead-of-larger-context/ (visited on 08/11/2023).

[14] Nils Reimers et al. *Say Goodbye to Irrelevant Search Results: Cohere Rerank Is Here*. en. May 2023. URL: https://txt.cohere.com/rerank/ (visited on 09/11/2023).

[15] Hongjin Su et al. *One Embedder, Any Task: Instruction-Finetuned Text Embeddings*. arXiv:2212.09741 [cs]. May 2023. DOI: 10.48550/arXiv.2212.09741. URL: http://arxiv.org/abs/2212.09741 (visited on 09/11/2023).

[16] Eduardo M. Moreira et al. "Prevalence of Pulmonary Hypertension in the General Population: The Rotterdam Study". eng. In: *PloS One* 10.6 (2015), e0130072. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130072.

[17] Gérald Simonneau et al. "Haemodynamic definitions and updated clinical classification of pulmonary hypertension". en. In: *European Respiratory Journal* 53.1 (Jan. 2019). Publisher: European Respiratory Society Section: Series. ISSN: 0903-1936, 1399-3003. DOI: 10.1183/13993003.01913-2018. URL: https://erj.ersjournals.com/content/53/1/1801913 (visited on 10/02/2023).

[18] Aakanksha Chowdhery et al. *PaLM: Scaling Language Modeling with Pathways*. arXiv:2204.02311 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2204.02311. URL: http://arxiv.org/abs/2204.02311 (visited on 09/11/2023).

[19] Rohan Anil et al. *PaLM 2 Technical Report*. arXiv:2305.10403 [cs]. May 2023. DOI: 10.48550/arXiv.2305.10403. URL: http://arxiv.org/abs/2305.10403 (visited on 09/11/2023).

[20] Jason Wei et al. *Emergent Abilities of Large Language Models*. arXiv:2206.07682 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2206.07682. URL: http://arxiv.org/abs/2206.07682 (visited on 09/11/2023).

[21] Takeshi Kojima et al. *Large Language Models are Zero-Shot Reasoners*. arXiv:2205.11916 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2205.11916. URL: http://arxiv.org/abs/2205.11916 (visited on 09/11/2023).

[22] Myriam Amsallem et al. "Addressing the Controversy of Estimating Pulmonary Arterial Pressure by Echocardiography". eng. In: *Journal of the American Society of Echocardiography: Official Publication of the American Society of Echocardiography* 29.2 (Feb. 2016), pp. 93–102. ISSN: 1097-6795. DOI: 10.1016/j.echo.2015.11.001.

[23] Jonathan D. Rich et al. "Inaccuracy of Doppler echocardiographic estimates of pulmonary artery pressures in patients with pulmonary hypertension: implications for clinical practice". eng. In: *Chest* 139.5 (May 2011), pp. 988–993. ISSN: 1931-3543. DOI: 10.1378/chest.10-1269.

[24] Micah R. Fisher et al. "Accuracy of Doppler echocardiography in the hemodynamic assessment of pulmonary hypertension". eng. In: *American Journal of Respiratory and Critical Care Medicine* 179.7 (Apr. 2009), pp. 615–621. ISSN: 1535-4970. DOI: 10.1164/rccm.200811-1691OC.

[25] Yongchun Shen et al. "CT-Base Pulmonary Artery Measurementin the Detection of Pulmonary Hypertension". In: *Medicine* 93.27 (Dec. 2014), e256. ISSN: 0025-7974. DOI: 10.1097/MD.0000000000000256. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4602811/ (visited on 09/11/2023).

A Appendix A

A.1 Regular Expression (Regex) Retrieval

We present below the regular expression rules utilized to identify pertinent text snippets for diagnosing pulmonary hypertension (PH) within the retrieval-augmented generation (RAG) approach.

```
r'(?i)(?:\bPulm.{0,10}hypertension\b|\bPH\b|\bPulm.?HTN
\b|\bp.?HTN\b|\bp.?AH\b|\barterial.hypertension\b|\bPHT\b|
(?:\belevated\b|\bhigh\b).(?:\bPASP\b|\bpulm.{0,10}art.
{0,5}sys|\bpulm.{0,10}art.{0,5}pr|\bPAP\b)|(?:\belevated\b|
\bhigh\b).(?:\bRVSP\b|\bRVP\b|\br.{0,5}v.{0,15}sys.{0,7}
pressure\b|\br.{0,5}v.{0,15}pressure\b)\b|\bflat.
{0,7}septum\b|\bseptal.flat|(?:\benlarge.{0,15}|
\bdilat.{0,15})\bpulm.{0,10}art|\bPH-ILD\b|\bPHILD\b|
\bCTEPH\b|\bPH-COPD\b|\bPHCOPD\b)')
```

These terms were curated by a physician SME using their domain knowledge. Future work could explore automated, or semi-automated, data mining methods to curate or expand relevant terms.

A.2 Pulmonary Hypertension Structured Phenotype

Code Vocabulary	Code	Name
ICD-9	416	Chronic pulmonary heart disease
	416.0	Primary pulmonary hypertension
	416.1	Kyphoscoliotic heart disease
	416.2	Chronic pulmonary embolism
	416.8	Other chronic pulmonary heart diseases
	416.9	Chronic pulmonary heart disease, unspecified
ICD-10	I27.21	Secondary pulmonary arterial hypertension
	I27.22	Pulmonary hypertension due to left heart disease
	I27.23	Pulmonary hypertension due to lung diseases and hypoxia
	I27.24	Chronic thromboembolic pulmonary hypertension
	I27.29	Other secondary pulmonary hypertension
	I27.8	Other specified pulmonary heart diseases
	I27.89	Other specified pulmonary heart diseases
	I27.9	Pulmonary heart disease, unspecified
RxNorm	1439816	Ambrisentan
	1439816	Bosentan
	8814	Epoprostenol
	40138	Iloprost
	1442132	Macitentan
	1439816	Riociguat
	1729002	Selexipag
	136411	Sildenafil
	358263	Tadalafil
	343048	Treprostinil

Table 2: Diagnostic and medication codes that make up the structured phenotype for PH.

A.3 Validation Experimental Results

This section presents the phenotyping results of the LLM-based architecture for the various configurations assessed in the validation dataset (Table 3) and enumerates the different types of prompt strategies that were considered (Table 4).

Prompt	LLM - Same Prompt			LLM - Different Prompt			Max Aggregation		
	No excl.	Regex excl.	Prompt excl.	No excl.	Regex excl.	Prompt excl.	No excl.	Regex excl.	Prompt excl.
A	0.67	0.69	**0.82**	0.71	0.72	0.64	0.78	**0.83**	**0.80**
B	0.58	0.56	–	0.56	0.60	–	0.73	0.69	–
C	0.56	0.55	–	0.68	0.71	–	0.58	0.55	–
D	0.74	0.77	–	0.74	0.74	–	0.74	0.77	–
E	0.76	0.77	–	0.59	0.64	–	0.78	0.83	–
Average	0.66	0.67	0.82	0.66	0.68	0.64	0.72	0.73	0.80

Table 3: Performance of different permutations of the LLM-based phenotype pipeline in the validation set, measured in terms of F_1 score. We conducted iterations using five different prompt designs and employed three distinct aggregation methods in the MapReduce stage (see the Aggregation section in the paper). Furthermore, we conducted experiments involving the removal of echocardiogram (ECHO) and computerized tomography (CT) reports, both through regular expression mentions and by appending exclusion instructions to the LLM prompt. The results in red indicate the configurations selected for comparison with the structured phenotype in the test set.

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

Table 4: Descriptions of the various prompt templates explored during the evaluation. Steering column refers to steering the LLM to count history of PH as a "no" and possibly having PH as a "yes". (*) In template C, prompt was modified to remove the phrase "explain your reasoning"; which is included in all other prompt templates.

A.4 Test Set Retrieval Note Type Distribution

Note Type	Frequency
Progress Note	54.21%
Consult	7.95%
Discharge Summary	6.06%
History & Physical Exam	5.77%
Procedure	5.53%
Telephone Encounter	3.94%
Other	16.54%

Table 5: Distribution of note types retrieved via Regex that are input into the LLM within the test set. "Other" consists of 23 different note types.