PhenoRAG: Retrieval-Augmented Generation for Efficient Zero-Shot Clinical Phenotype Identification

Marc Berndt MBERNDT@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Zurich, Switzerland

Andrea Agostini
Department of Computer Science, ETH Zurich, Zurich, Switzerland

Beatrice Stocker
Division of Metabolism and Children's Research Center, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland

Maria Padrutt
Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

Silvio D Brugger
Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

D Sean Froese
Division of Metabolism and Children's Research Center, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland

Daphné Chopard* DAPHNE.CHOPARD@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Zurich, Switzerland; Department of Intensive Care and Neonatology, University of Zurich, University Children's Hospital Zürich, Zurich, Switzerland

Julia E Vogt* JULIA.VOGT@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Zurich, Switzerland

Abstract

Accurate extraction of phenotypic information from clinical narratives is essential in diagnostic medicine, yet mapping free-text reports to structured Human Phenotype Ontology (HPO) terms remains challenging. While encoder-only transformer models and small decoder-only generative models are attractive for clinical deployment due to their efficiency and low resource requirements, the former often fail to capture the rich context of clinical texts, and the latter struggle to process lengthy reports effectively. In contrast, larger language models excel at contextual understanding but are impractical for clinical use due to their size, propensity to hallucinate, and privacy concerns associated with non-local inference. To overcome these challenges, we introduce PhenoRAG, a novel retrieval-augmented generation framework that leverages a synthetic database of contextually enriched sentences to augment a lightweight decoder-only model for accurate zero-shot phenotype identification. We demonstrate the capacity of PhenoRAG to capture nuanced contextual clues by 1) evaluating its ability to perform two clinically relevant tasks—guide rare disease diagnosis and facilitate urinary tract infection detection—and 2) validating its performance on a synthetic dataset designed to mimic the challenges of real clinical narratives. Experimental results demonstrate that our lightweight PhenoRAG framework achieves a higher F1-score than both encoder-only transformers and standalone small lan-

* Equal contribution

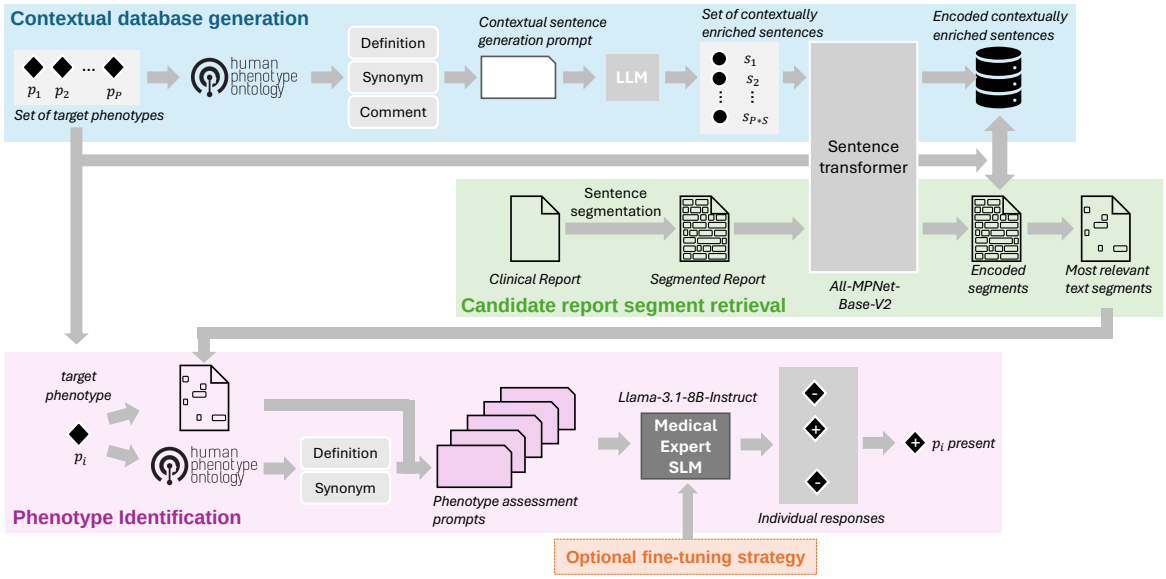

Figure 1: Schematic of PhenoRAG for phenotype identification. For each phenotype, S relevant clinical report segments are retrieved and combined with shared HPO-derived information about p_i to create S augmented prompts. Each prompt is processed by an SLM that independently evaluates the presence of the phenotype. The individual predictions are then aggregated into a final decision. An optional fine-tuning module can be incorporated to tailor the model's performance.

guage models, driven primarily by its high recall. These findings underscore the potential of PhenoRAG as a ready-to-use clinical tool for phenotype identification[1].

1. Introduction

Extracting phenotypes from unstructured clinical narratives is crucial for healthcare applications as it can enable precise diagnoses, patient stratification, and tailored therapeutic interventions. A widely adopted strategy is to leverage the Human Phenotype Ontology (HPO) (Köhler et al., 2021), a hierarchical framework in which each phenotype is linked to a unique code. However, reliably mapping free-text clinical reports to HPO terms remains challenging due to the heterogeneity, complexity, and nuanced language found in real-world clinical texts.

Encoder-only transformer models, such as Phenotagger (Luo et al., 2021) and PhenoBERT (Feng et al., 2022), have emerged as competitive low-resource solutions suitable for clinical deployment. However, their limited contextual understanding restricts their effectiveness when faced with indirect phenotype descriptions, family histories, temporal references, or negations common in clinical practice. In contrast, large decoder-only generative language models (LLMs) excel in capturing contextual subtleties and semantic complexities but are often impractical for clinical use due to their large size, high computational demands, and tendency to hallucinate when used off-the-shelf (Yang et al., 2024; Groza

1. The code is publicly available: https://github.com/marcb1893/PhenoRAG

et al., 2024; Stinton et al., 2023; Wang et al., 2024). Furthermore, privacy concerns necessitate local inference, which further limits the applicability of these models. Smaller variants, while more feasible for local deployment, struggle to effectively process lengthy clinical documents (Groza et al., 2024).

To overcome these challenges, we introduce PhenoRAG, a retrieval-augmented generation (RAG) framework for accurate phenotype identification in low-resource settings (see Figure 1). PhenoRAG leverages a synthetic database of contextually enriched sentences generated from detailed HPO descriptions. These sentences are encoded to retrieve the most pertinent segments from clinical reports, and the resulting text fragments augment a small decoder-only language model (SLM) for phenotype identification. Although PhenoRAG operates effectively in a zero-shot setting, optional fine-tuning with carefully constructed synthetic data further enhances its ability to distinguish closely related concepts, handle typographical errors, and correctly interpret negations.

We evaluate PhenoRAG on two real-world clinical datasets as well as a synthetically generated dataset designed to replicate the challenges of clinical narratives. Our experimental results demonstrate that PhenoRAG achieves significantly higher recall and overall performance compared to existing encoder-only transformer models, including PhenoBERT. Moreover, its small size, readiness to be used—with no fine-tuning required—and suitability for local inference, underscore its potential for practical deployment in clinical settings.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work reveals the importance of aligning algorithmic performance with actual clinical requirements in the context of extracting clinical data. This includes the need to filter irrelevant phenotypes and have sensitivity to negations. We demonstrate that current extraction methods, developed on such standardized datasets, often overlook these nuances, resulting in poor real-world performance. Our novel method, tested on real world datasets, not only addresses these limitations, but is also usable for deployment in typical healthcare environments.

2. Related Work

CNN-based approaches Early approaches such as NeuralCR (Arbabi et al., 2019) used a CNN-based architecture to encode clinical text and map it to HPO codes via an ancestry-informed embedding matrix, effectively capturing hierarchical relations.

BERT-based approaches PhenoTagger (Luo et al., 2021) advanced the capabilities of NeuralCR by integrating a Trie tree-based dictionary search for exact matches with a transformer-based tagger (using BioBERT (Lee et al., 2020) as the backbone) to handle spelling variants and ambiguity—albeit with increased computational overhead. Building on these models, PhenoBERT (Feng et al., 2022) combined the efficiency of CNN-based encoding with BERT's accuracy through a two-level hierarchical approach that narrows candidate concepts before applying fine-grained classification. Although these BERT-based solutions are targeted and lightweight, they generally fall short in capturing and interpreting complex contextual information, limiting their practical efficacy in clinical settings despite high performance on standard datasets.

LLM-based approaches Owing to their strong contextual understanding, generative approaches have also been explored for HPO concept recognition (Groza et al., 2024; Stinton et al., 2023; Yang et al., 2024). However, LLMs such as GPT-4 exhibit issues like hallucinations, making them ill-suited for precise identification (Stinton et al., 2023; Wang et al., 2024). Although in-context learning can partially mitigate this issue (Groza et al., 2024), challenges related to model size and the selection of appropriate context persist. These limitations hinder accurate local inference—particularly in settings where data privacy is critical.

RAG-based approaches Recent work has explored retrieval-augmented generation for phenotype identification. For example, RAG-HPO (Garcia et al., 2024) integrates retrieval with a large (70B parameter) LLM to improve HPO concept recognition, achieving strong recall but at a high computational cost that limits local deployment. Similarly, semantic similarity retrieval has been shown to enhance annotations on synthetic data (Albayrak et al., 2025). While direct retrieval-based annotation is infeasible for real-world data, the framework may still be able to provide key information for subsequent response generation.

In this work, we leverage a RAG framework with a SLM, thereby addressing the limitations of both traditional deep learning methods and large-scale LLMs while enabling efficient, real-world clinical application. Our novel approach addresses key limitations of existing LLMs—prone to hallucinations—and of SLMs that lack the capacity for such a complex task, by integrating a retrieval-augmented generation framework with a lightweight decoder-only language model for phenotype identification.

3. Methods

PhenoRAG is a retrieval-augmented generation framework that leverages a synthetic database of contextually enriched sentences—derived from detailed HPO descriptions—to extract the most relevant segments from lengthy clinical narratives and augment an SLM for zero-shot phenotype identification. The process begins by defining a target set of phenotypes to identify based on the task at hand. These are used to create a synthetic database of contextually enriched sentences describing each target phenotype. PhenoRAG then segments the clinical report into smaller sentence-like portions. For each target phenotype, the model compares the report segments with the corresponding sentences from the synthetic database and selects the most pertinent report segments, discarding the rest. The selected report segments are used—along with detailed information about the phenotype extracted from the HPO—to augment a prompt supplied to a "Medical Expert" SLM which is tasked to determine whether or not the report segment indicates the presence of the target phenotype. Finally, the individual responses of the "Medical Expert" SLM for each retrieved report segment are aggregated. Based on the aggregated outcome, the phenotype is determined to be present or absent. The complete approach is illustrated in Figure 1, with each component described in detail in the remainder of this section.

3.1. Contextual Database Generation

To effectively retrieve relevant clinical report segments, PhenoRAG begins by constructing a synthetic database consisting of contextually enriched sentences for each target pheno-

type. Specifically, for each target phenotype p_i ($i \in [1, P]$, with P being the total number of target phenotypes), we prompt a decoder-only generative language model to generate representative sentences that reflect the diverse contexts in which these phenotypes typically occur in clinical narratives. The language model is encouraged to incorporate varying clinical perspectives and utilize authentic clinical shorthand notation, following the prompt strategy proposed by Albayrak et al. (2025) (Figure 3 in Appendix A.1). Next, each synthetically generated sentence is encoded into a dense vector representation via a sentence transformer. The resulting set of embeddings constitutes our contextual database, which serves as the basis for efficiently identifying and retrieving clinical report segments most relevant to the presence of each target phenotype.

3.2. Candidate Report Segment Retrieval

Once the contextual database has been constructed, it is used to identify and retrieve the most relevant segments from each clinical report for phenotype assessment. In this step, each report is initially divided into shorter text segments, which are then encoded using the same sentence transformer utilized for encoding the contextual database. The encoded report segments are subsequently compared against the contextual database via cosine similarity. For each target phenotype, the S segments with the highest similarity scores are selected. This effectively compresses the original clinical report into $S \cdot P$ short segments that serve to augment the "Medical Expert" SLM for targeted phenotype identification.

3.3. Phenotype Identification

The "Medical Expert" SLM is configured to act as a medical professional tasked with determining the presence or absence of specific phenotypes by responding strictly with "Yes" or "No" through a *system prompt*. In this step, for each target phenotype p_i, PhenoRAG extracts detailed information about the phenotype—its name, definition, and synonyms—from the HPO. Each of the S retrieved report segments is individually combined with this phenotype-specific information to augment the SLM's prompt (*user prompt*), asking whether the segment indicates the presence of the given phenotype (see Figure 2). Finally, the individual S responses from the "Medical Expert" SLM are aggregated: a single "Yes" response indicates a positive classification for the phenotype, whereas the absence of any "Yes" responses indicates insufficient evidence for the phenotype's presence.

3.4. Efficient Fine-Tuning for Enhanced Symptom Differentiation

We introduce PhenoRAG-FT, a fine-tuned variant of PhenoRAG designed to enhance the SLM's performance in challenging or task-specific scenarios. While standard PhenoRAG already performs well without fine-tuning, this optional low-resource strategy can target specific difficulties—such as distinguishing closely related phenotypes, handling typographical errors, and interpreting complex negations. A synthetic dataset carefully crafted to mimic these challenges is generated and used to fine-tune PhenoRAG's "Medical Expert" SLM and thereby improve its contextual precision. This fine-tuning approach is flexible and can be adapted to other tasks by modifying the synthetic data accordingly. The two main steps are described below and more details can be found in Appendix A.2

System prompt:

"You are a medical expert deciding whether a patient has a certain symptom. Answer only with 'Yes' or 'No'."

User prompt:

"The symptom <Label> is defined as <Definition>.
<Label> is also referred to as <Synonyms>.
Does the following text segment explicitly confirm that the patient has this symptom: <Input >."

Figure 2: Prompt template for phenotype assessment. Expressions in angle brackets are replaced with the phenotype's name (<Label>), further details extracted from the HPO (<Definition> and <Synonyms>), and the report segment under evaluation (<Input>).

Synthetic Fine-tuning Dataset Creation To create a robust fine-tuning dataset, we start by using sentences from the context database as positive examples labeled "Yes." Negative examples are generated by mismatching phenotypes—embedding a sentence for phenotype p_j into the prompt for p_i (where $i \neq j$)—as well as by synthesizing hard negatives from HPO-neighboring concepts using a decoder-only language model. To improve real-world robustness, the dataset is further augmented with typographical errors and negated statements. Typo variants are created using a character-level mutation function applied to symptom names and synonyms. Negated examples are generated via predefined templates that express absence, while additional positive examples—both keyword-based and descriptive—are produced using a separate generator, with and without intentional spelling errors. Together, these steps ensure rich coverage of real-world clinical language and edge cases. More details are provided in Appendix A.2.

Fine-tuning PhenoRAG-FT is fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2022), which inserts lightweight trainable matrices into the attention layers of the transformer, allowing efficient parameter updates with minimal resource requirements. The model is optimized using a standard causal language modeling objective (cross-entropy loss), enabling domain-specific adaptation without full model retraining. This setup ensures efficient, scalable fine-tuning suitable for local clinical deployment.

4. Automated Phenotype Extraction for Rare Disease Diagnosis

Rare genetic disorders pose significant diagnostic challenges, often requiring months or even years for a definitive diagnosis (Saudubray et al., 2022). In addition, due to their low prevalence, patients are distributed across multiple centres and countries, making it difficult to compare clinical narratives. Yet, phenotypic information is crucial for diagnosing these conditions—especially when integrated with genomic and other omics data—and to describe the natural history of disease, which is invaluable to assess diagnostic screening tools. An automated, standardized approach to processing clinical reports can therefore help compare

patient profiles and subtype disorders, with the ultimate goal of identifying disease-relevant phenotypes for diagnosis.

In this study, our method is applied to a cohort of patients affected by remethylation disorders. Remethylation disorders are a group of rare inherited metabolic conditions caused by defects in the pathway that converts homocysteine to methionine, resulting in elevated plasma homocysteine and a broad spectrum of clinical manifestations. By automatically extracting key phenotypes from heterogeneous clinical reports, the approach provides deeper insights into the patients' clinical presentations and underlying pathologies, thereby supporting more effective disease subtyping and diagnosis.

4.1. Remethylation Disorder Dataset

Overview This dataset consists of 118 clinical reports collected over 30 years by the University Children's Hospital Zurich, where individuals from all over Europe were referred to with suspicion of a remethylation disorder (RD). All individuals presented with elevated plasma homocysteine levels. In 44 individuals this is the only informative diagnostic metabolite identified (therefore termed: isolated hyperhomocysteinemia) while 74 individuals additionally have increased urinary methylmalonic acid (termed: combined hyperhomocysteinemia and methylmalonic aciduria). Among the 44 individuals with isolated hyperhomocysteinemia, 22 have severe MTHFR deficiency (confirmed via identification of bi-allelic pathogenic variants in the *MTHFR* gene), while 52 of the 74 individuals with combined hyperhomocysteinemia and methylmalonic aciduria have cbl$C$ deficiency (confirmed via identification of homozygous c.394C>T ($n = 26$) or c.271dupA ($n = 26$) variants in the *MMACHC* gene). The causal variant(s) for the remainder of the cohort is unknown. A cohort summary is available in Table 7 in Appendix D.1. Originally provided as scanned PDFs, the reports were digitalized and manually curated for phenotype-relevant sections. When necessary, they were translated into English using a deep-learning translator[2]. Ground-truth annotations were rigorously established and reviewed by at least two annotators in consultation with medical experts.

Challenges The clinical documents are highly heterogeneous, encompassing all patient-relevant records collected by the hospital—such as email correspondence with referring physicians, follow-up communications, enzymatic assay results, and genetic test reports. This diversity leads to wide variations in document length, reporting style, and the level of detail provided, reflecting differences among referring physicians and institutions. Consequently, the heterogeneity in how phenotypes are documented calls for a flexible approach that does not rely solely on fixed textual patterns. Furthermore, the indirect description of symptoms (e.g., *"at 5 months [...] developmental age was about 2 months"* instead of the explicit phenotype "global developmental delay") and the common interweave of detailed family histories—which is paramount for rare disease diagnosis but should be ignored in this task—poses significant challenges for current text-to-HPO mapping methods. Notably, of the 118 reports, 100 included detailed clinical phenotypes, while the remainder contained only treatment and biochemical data. Including these reports in the evaluation further challenges the approach to ignore irrelevant information.

2. Original reports were written either in German, English, French, or Italian

4.2. Experiment

Based on the clinical guidelines for diagnosis and treatment of remethylation disorders (Huemer et al., 2017), we focus on identifying the following nine phenotypes: feeding difficulties, failure to thrive, seizures, abnormal muscle tone (hypotonia/hypertonia), neurodevelopmental delay (NDD), anemia, thrombocytopenia, and cerebral atrophy. Tables 8 and 9 in Appendix D.1 provide an overview of the target phenotypes, their HPO Code, their occurrence in the dataset and a short description. Note that 79 of the total 118 reports contained at least one of the assessed phenotypes or a more specific term of these. If a more specific phenotype was present (e.g., "motor delay" instead of NDD), it was mapped to the more general term.

Experiment Details To evaluate PhenoRAG, we apply the pipeline as described in Section 3 above. First, the contextual database is created by prompting the *Llama-3.1-70B-Instruct* model (Grattafiori et al., 2024) to generate 40 synthetic sentences for each phenotype. These sentences are encoded with the *All-MPNet-Base-V2* sentence transformer (Song et al., 2020) and will serve as a comparison for the retrieval of the relevant report segments. Then, each clinical report is preprocessed using Stanza for deep learning-based segmentation (Qi et al., 2020; Zhang et al., 2021). To avoid excessively long sections, the segmented text is further divided at newline characters, and each resulting fragment is encoded once again with the *All-MPNet-Base-V2* sentence transformer (Song et al., 2020). The number of retrieved segments per target symptom is set to $S = 5$, following prior work Albayrak et al. (2025). Finally, as the Medical Expert for phenotype identification, we use the SLM *Llama-3.1-8B-Instruct* (Grattafiori et al., 2024) in 8-bit quantization with scaled dot-product attention. Sampling is disabled for deterministic output. Given the limited dataset size and the zero-shot nature of our approach, we did not split the dataset into separate validation and test sets. Instead, all reported results are evaluated on the full set of clinical reports. To avoid overfitting and maintain the integrity of the evaluation, we also refrained from extensive hyperparameter tuning and relied on standard, previously established values.

Benchmarks We compare our approach, PhenoRAG, with a state-of-the-art BERT-based method for phenotype identification, PhenoBERT[3]. BERT-based models are attractive for clinical applications due to their lightweight nature, local inference capabilities, and suitability for environments with strict privacy requirements. However, preliminary experiments revealed that PhenoBERT struggles with negation detection. To address this, we enhanced PhenoBERT by integrating the Clinical Assertion and Negation Classification BERT (CANBERT) model (Van Aken et al., 2021; Su et al., 2024) (see Appendix B.2 for full details), and we evaluate this improved variant alongside the original.

4.3. Results and discussion

Table 1 summarizes the evaluation metrics, reported as both micro and macro averages in line with previous work (Feng et al., 2022). The micro average aggregates predictions across all phenotypes regardless of their report of origin, while the macro average scores are computed on a per-report (i.e. per-patient) basis (additional details are provided in

3. Appendix B.1 provides a comparison of the existing BERT-based approaches on a standard dataset

Appendix C). Note that 29 of the 118 reports did not contain any target symptoms; for these reports, precision and recall are conventionally set to 1 or 0, which can disproportionately influence the macro-average metrics and should therefore be interpreted with caution.

Table 1: Performance comparison of PhenoRAG versus PhenoBERT on the RD dataset.

Model	**Micro Average [%]**			**Macro Average [%]**		
	P	**R**	**F1**	**P**	**R**	**F1**
PhenoBERT	94.64	74.30	83.25	90.96	81.46	85.95
PhenoBERT+CAN-BERT	**97.55**	74.30	84.35	**92.30**	81.68	86.66
PhenoRAG	89.81	**90.65**	**90.23**	89.08	**89.88**	**89.48**

Overall, PhenoRAG achieves a significantly higher F1-score—most notably in the micro-average metrics, with an almost 6-percentage-point increase. This improvement is primarily driven by a substantial increase in recall (+16.35 percentage points), despite a slight decrease in precision (-4.83 percentage points compared to PhenoBERT and -7.74 percentage points relative to PhenoBERT+CAN-BERT). Although the macro-average metrics are somewhat biased by reports that lack any phenotype, PhenoRAG still exhibits a higher F1-score (between +2.82 and +3.53 percentage points), again largely due to its impressive recall, even with slightly lower precision. These results underscore the clinical potential of PhenoRAG: in medical applications, high recall (or sensitivity) is crucial to ensure that no important phenotype is overlooked, and a slight reduction in precision can be acceptable, especially when it can be compensated for by a quick, low-effort verification of the key text segments.

Table 13 in Appendix E.1 provides insight into the error profiles of the evaluated models and highlights why PhenoRAG achieves superior recall. First, PhenoRAG correctly interprets simple negations directly, whereas PhenoBERT requires the integration of CAN-BERT to handle such cases. More impressively, PhenoRAG is able to pick up on complex negations that require a broader context understanding, an area where PhenoBERT struggles even equipped with state-of-the-art clinical negation detection. For example, in the sentence "mum reports a developmental delay initially but this has picked up in the last two months", PhenoRAG correctly detects temporal cues that effectively invalidate the phenotype, thus avoiding a false-positive prediction, unlike the BERT-based models. Furthermore, PhenoRAG accurately interprets complex narrative descriptions—for example, recognizing "can sit with support only, never walked or acquired speech" as indicative of Neurodevelopmental Delay, and "height-related body weight of 9.2 kg was below the normal range (<P3)" as a signal for Failure to Thrive. This nuanced understanding enables PhenoRAG to capture clinically relevant information that simpler, BERT-based approaches miss. Moreover, PhenoRAG effectively discards phenotypes related to family history, addressing a key limitation of BERT-based models that struggle with contextual understanding. However, this capability depends on the text segment retrieval component capturing sufficient context. For example, the segment "One was a male infant with psychomotor retardation", following a reference to the patient's siblings, lacked enough context for the SLM to infer that it did not apply to the patient, leading to a false positive. However, such challenging instances—which are also missed by BERT-based models—could be mitigated in PhenoRAG by retrieving larger text segments. Occasionally PhenoRAG misinterprets

descriptions resulting in a drop in precision. For example, the phrase "Intra-uterine growth delay" is misclassified by PhenoRAG as Failure to Thrive, probably because by the Failure to Thrive's HPO entry includes as a synonym "undergrowth", even though intra-uterine growth delay (IUGD, HP:0001511) is a distinct phenotype in the ontology. We speculate that if IUGD were included in the target symptoms, PhenoRAG would have identified that phenotype correctly, even though this would not have an impact on the erroneous prediction for Failure to Thrive. Nevertheless, in the next experiments, we demonstrate that fine-tuning the SLM model to effectively differentiate between neighbouring phenotypes can effectively reduce these errors. Overall, PhenoRAG demonstrates a robust ability to interpret extended and complex contexts—including subtle temporal cues and indirect phenotype assertions—that are common in clinical texts (see further examples in Table 13). It is important to recall that these results are achieved without any task-specific fine-tuning or prior exposure to the data before inference.

5. Bacteriuria Classification for Antibiotic Use Reduction

Antimicrobial resistance (AMR) is a leading global health threat, accounting for an estimated 1.27 million deaths in 2019 and projected to reach up to 10 million deaths annually by 2025 if unmitigated (Thangaraju and Venkatesan, 2019; Salam et al., 2023; Murray et al., 2022; O'Neill, 2014). Excessive and inappropriate antibiotic use is the main driver of AMR (Salam et al., 2023; Sanchez et al., 2016; Swami et al., 2012; Palmer and Kishony, 2013). In this context, urinary tract infections (UTIs) represent 15–21% of all antibiotic prescriptions yet are frequently overdiagnosed (Timm et al., 2024; Aabenhus et al., 2017). In clinical practice, antibiotic prescription is often based solely on the detection of bacteria in urine (bacteriuria), even though treatment is unwarranted in cases of asymptomatic bacteriuria (ASB) (Nicolle et al., 2005). To address this, accurately distinguishing UTI from ASB is critical.

In this section, phenotype identification is applied to unstructured clinical reports from hospitalized individuals with reported bacteriuria. By automatically detecting UTI-specific symptoms—or the lack thereof—potential ASB cases can be flagged and unnecessary antibiotic treatment can be stopped earlier. We evaluate our approach on a real-world hospital dataset and further test it using a synthetic dataset that captures challenging, descriptive clinical narratives. This dual evaluation allows us to compare our method against lightweight, clinically suitable benchmarks as well as a larger model, confirming the potential and competitiveness of PhenoRAG for clinical deployment.

5.1. Hospital UTI Dataset

Overview This private dataset originates from the University Hospital Zurich and consists of 30 clinical reports written in German documenting the hospital stays of nine patients with confirmed bacteriuria, representing nine cases of suspected urinary tract infections (UTIs). The reports encompass a wide range of clinical documentation collected during the hospital stay, including admission notes, progress updates, and transfer summaries. A clinical expert reviewed each case and categorized them into one of two groups based on established guidelines (Cortes-Penfield et al., 2017): six individuals exhibited UTI symp-

toms (including uncomplicated, complicated, or catheter-associated UTI) and were deemed eligible for antibiotics, while three individuals were identified as having ASB.

Pre-processing As a pre-processing step, the original German reports were translated into English using an offline solution[4] to ensure data privacy. All reports related to each case were then concatenated into a single comprehensive document. These aggregated documents averaged 21,000 characters (approximately 3300 words) each. Additional details about the dataset are provided in Appendix D.2.

Challenges A major challenge in this task is extracting the few UTI-specific phenotypes from lengthy and complex clinical reports, which often include a large number of unrelated symptoms. Because bacteriuria is rarely the primary reason for hospitalization, these hospital reports are typically saturated with clinical details irrelevant to UTI detection. Longer documents also increase the risk of false positives, as models are more likely to encounter ambiguous phrases or loosely matching terms that either resemble the target phenotype or correspond to a different one entirely. Additionally, varying formatting and reporting styles across departments further complicate inference, especially for models optimized on more standardized datasets. Together, these factors make this task particularly challenging for most phenotype identification methods.

5.2. Experiment 1: UTI detection from hospitalized patient records

This first experiment focuses on UTI-specific phenotype identification in hospital clinical reports to facilitate the detection of UTI cases. When bateruria is present, a UTI diagnosis is only confirmed if UTI-specific phenotypes are also reported. In the absence of such symptoms, the condition is classified as asymptomatic bacteriuria (ASB), which does not require antibiotic treatment. Table 2 lists the UTI-specific phenotypes used in this task to distinguish between UTI and ASB cases. These phenotypes were selected by infectious disease experts in accordance with established clinical guidelines and domain expertise (Cortes-Penfield et al., 2017). For each patient, we manually reviewed the clinical reports to determine the presence or absence of these target phenotypes. This phenotype-level annotation provides an additional layer of ground truth—beyond the binary UTI vs. ASB label—enabling a more fine-grained comparison of PhenoRAG's performance against alternative approaches.

Experiment Setup Each patient case is represented by a concatenated clinical document, which is scanned for the presence of predefined UTI-specific phenotypes (see Table 2; full details in Appendix D.2). For each case, the phenotype extraction method must determine which, if any, of the target phenotypes are present. If at least one UTI-related phenotype is detected, the case is classified as a UTI. Conversely, if none of the target phenotypes are found, the case is labeled as asymptomatic bacteriuria (ASB). The pipeline is illustrated in Figure 7 in Appendix D.2.

Experiment Details Similar to the previous experiment (Section 4.2), for PhenoRAG, clinical reports are segmented using Stanza (Qi et al., 2020; Zhang et al., 2021) *All-MPNet-Base-V2* is used as a sentence transformer (Song et al., 2020) and we use *Llama-3.1-8B-*

4. PRoMT Master NMT 23 Multilingual

Instruct for the "Medical Expert" SLM (Grattafiori et al., 2024). As part of the contextual database generation, *Llama-3.1-70B-Instruct* (Grattafiori et al., 2024) is prompted to generate 40 synthetic sentences for each of the 5 target UTI phenotypes. We set $S = 5$ as retrieval cut-off. For fine-tuning, the Adam optimizer (Kingma and Ba, 2014) is used with a learning rate of $3 \cdot 10^{-4}$. The LoRA parameters are set to $r = 8$ and $\alpha = 32$. Regularization is applied with a dropout rate of 1%. Fine-tuning is performed for one epoch on an NVIDIA RTX 4090 GPU. The other default hyperparameters are retained to confirm performance stems from design, not tuning.

Given the importance of accurately distinguishing between closely related phenotypes—particularly for avoiding false positives in ASB cases—we also evaluate the fine-tuned variant of our method, PhenoRAG-FT. The fine-tuning procedure is described in Section 3.4 with further details in Appendix A.2. To generate the fine-tuning dataset, we use the large decoder-only generative model *Llama-3.3-70B-Instruct* (Grattafiori et al., 2024).

Table 2: Targeted phenotypes for UTI identification and their frequency in the dataset.

Phenotype	**HPO Code**	**Frequency** [%]
Dysuria	HP:0100518	44.4
Pollakisuria	HP:0100515	33.3
Flank pain	HP:0030157	33.3
Lethargy	HP:0001254	11.1
Chills	HP:0025143	0

Benchmarks To assess PhenoRAG's performance relative to state-of-the-art models suitable for hospital deployment, we compare it to two strong encoder-only baselines: PhenoBERT, a competitive representative of this class of models (see Appendix B.1), and a negation-aware extension, PhenoBERT+CAN-BERT (described in Section 4.2).

Additionally, we include as baseline the SLM (*Llama-3.1-8B*) itself without the retrieval component but with the same system prompt to turn it into the same "Medical Expert". This configuration serves two purposes: it evaluates the feasibility of using SLMs as standalone tools for phenotype identification—appealing in clinical practice due to their small size and local deployability—and acts as an ablation study to isolate the contribution of PhenoRAG's retrieval module.

Results and Discussion The classification performance for distinguishing UTI from ASB is presented in Table 3. Among the nine cases, both PhenoRAG and its fine-tuned variant PhenoRAG-FT misclassified only a single case, failing to identify UTI-specific symptoms in one instance. Notably, the high performance of PhenoRAG was achieved entirely out-of-the-box: neither the sentence encoder nor the SLM were fine-tuned on this task.

In contrast, baseline models showed more frequent errors: PhenoBERT misclassified five cases, PhenoBERT+CAN-BERT two, and the standalone *Llama-3.1-8B* model without the retrieval component three. These results underscore PhenoRAG's strong performance relative to both encoder-only transformers and an unstructured application of an SLM.

A closer examination of the predictions highlights a key limitation of using *Llama-3.1-8B* without retrieval: it detected UTI-specific symptoms in every report, regardless

of the actual content. As a result, its performance appears inflated due to the dataset's imbalance, where two-thirds of cases are UTI-positive. Its failure to distinguish between symptomatic and asymptomatic cases highlights the critical importance of the retrieval component in PhenoRAG, as SLMs alone have too limited capacity for such a complex task. This issue is further illustrated in the fine-grained phenotype-level evaluation (Table 4). Although the recall of the full-document *Llama-3.1-8B* baseline matches that of PhenoRAG and PhenoRAG-FT, its precision drops dramatically to 32.26%, compared to 90.91% for PhenoRAG-FT—demonstrating a strong tendency toward overprediction when relevant context is not isolated in advance.

Meanwhile, the comparison between PhenoBERT and PhenoBERT+CAN-BERT highlights the importance of clinical negation detection. The addition of CAN-BERT improves precision by 23.33 percentage points, confirming that negated symptoms are frequent in this dataset. In contrast, PhenoRAG naturally infers the absence of the phenotype *Chills* from phrases like "no fever or chills" without requiring additional enhancements. However, both models lag significantly behind the decoder-based approaches in recall (27.27 percentage points). Error analysis shows that the higher recall achieved by PhenoRAG is driven by its superior contextual understanding, enabling it to capture semantic nuances beyond simple keyword matching. For example, PhenoRAG accurately determines that the phrase "anamnestic regular micturition" does not indicate pollakisuria—whereas PhenoBERT erroneously assigned this symptom. This, once again indicates the enhanced ability of PhenoRAG to detect contextually complex or indirectly phrased symptoms. Nevertheless, PhenoRAG occasionally produces silent errors, such as erroneously confirming *Flank Pain* when encountering "ureterolithiasis". Targeted fine-tuning however helped reduce these errors by refining the model's conceptual representations as shown by PhenoRAG-FT's improved precision.

Although the retrieval system component of PhenoRAG drastically reduced the amount of text that was given to the SLM for phenotype assessment—from an average of 21,000 characters per report to just 1,250 characters (approximately 6% of the original size)—PhenoRAG and PhenoRAG-FT still achieved outstanding results, misclassifying only three and two phenotypes respectively across all reports. This confirms that the retrieval module effectively preserved clinically relevant information while discarding noise, enabling efficient and accurate assessment.

Table 3: Comparison on the hospital UTI dataset for **UTI case classification**. Values are in [%].

	Accuracy	**Precision**	**Recall**	**F1-Score**
PhenoBERT	55.55	62.50	83.33	71.43
PhenoBERT+CAN-BERT	77.77	83.33	83.33	83.33
Llama-3.1-8B	66.67	66.67	**100.00**	80.00
PhenoRAG	**88.88**	**100.00**	83.33	**90.89**
PhenoRAG-FT	**88.88**	**100.00**	83.33	**90.89**

Table 4: Comparison on the hospital UTI dataset for UTI **phenotype identification**. Values are in [%].

Model	Micro Average [%]			Macro Average [%]		
	P	**R**	**F1**	**P**	**R**	**F1**
PhenoBERT	46.67	63.64	53.85	35.33	43.33	37.86
PhenoBERT+CAN-BERT	70.00	63.64	66.67	45.33	43.44	62.43
Llama-3.1-8B	32.26	**90.91**	47.62	30.36	**75.00**	41.24
PhenoRAG	83.33	**90.91**	86.96	69.33	73.33	71.11
PhenoRAG-FT	**90.91**	**90.91**	**90.91**	**73.33**	73.33	**73.33**

5.3. Synthetic UTI Dataset

Overview For the second dataset, we generated a set of 30 synthetic reports using different LLMs with the prompt provided in Figure 8 of Appendix D.3. This prompt was designed to include statements about the targeted phenotypes[5] as well as to replicate the linguistic challenges found in real clinical documents by incorporating stylistic variations such as negations and paraphrases. We used GPT-4o (Achiam et al., 2023), GPT-4o-mini (Hurst et al., 2024), Claude 3.5 Sonnet (Anthropic, 2024), and Claude 3.7 Sonnet (Anthropic, 2025) as LLMs to ensure some variability between the generated reports as the same model tends to generate similar outputs after a while. As part of the generation, each LLM was prompted to output a list of phenotypes intended to be present in the report. These annotations were manually reviewed to ensure they aligned with the HPO definitions and no neighbouring concepts were more appropriate. As a result, each document was precisely mapped to specific HPO Codes, which were then used as ground truth. More details about the dataset can be found in Tables 11 and 12 in Appendix D.3.

Challenges This dataset was intentionally designed to embed symptoms within complex, descriptive language, posing a strong challenge for phenotype identification methods. It requires models to truly understand contextual meaning rather than relying on keyword matching or surface-level cues.

5.4. Experiment 2: Synthetic Dataset Evaluation

In this second experiment, the goal is to determine whether each of the reports in the synthetic UTI dataset contains any of the five UTI-specific target phenotypes. Each report is scanned independently for the presence or absence of these phenotypes, allowing a fine-grained assessment of model performance under controlled yet challenging conditions.

Benchmarks In addition to the clinically suitable baselines introduced in Section 4.2, we include an additional comparison in this experiment: a large-scale RAG-based method tailored for LLMs: RAG-HPO (Garcia et al., 2024). Unlike the private clinical dataset, this

5. Based on the idea of reducing antibiotic usage and differentiating between ASB and UTI, we use the symptoms in Table 2 to generate this dataset.

synthetic dataset allows for evaluation in a more resource-rich computational environment, making it feasible to compare PhenoRAG against larger models.

Results and Discussion In Table 5, we report both the micro-average (phenotype-level) and macro-average (document-level) metrics for phenotype identification (see Appendix C). On the synthetic dataset, PhenoBERT models have a very low recall of about one third of the target phenotypes (35.21%), while the RAG-based models achieved recall levels from 80% to over 98%. This strong difference stems from the complexity of symptom descriptions, a strength of PhenoRAG but which BERT models struggle to capture. Additionally, PhenoBERT's precision was lower due to issues such as undetected complex negations and erroneous mappings, as detailed in Table 14. These findings indicate that the inherent limitations of PhenoBERT are likely to become even more pronounced with larger, more heterogeneous datasets. In contrast, the larger decoder-only LLM within the RAG framework (RAG-HPO) better interprets contextual cues—effectively differentiating between current and past conditions—resulting in improved precision. However, RAG-HPO relies on a 70B parameter model, whereas PhenoRAG achieves competitive precision, especially its fine-tuned variant, when relying only on an 8B parameter model.

The vanilla PhenoRAG model achieves higher recall, while fine-tuning significantly boosts precision (see Table 5). More specifically, it improves its ability to distinguish between closely related symptoms, handle complex negations, and reject misleading causal or co-occurring conditions, thus enhancing precision and reducing silent errors. However, this increased specificity sometimes comes at the cost of recall for borderline cases, making the choice between PhenoRAG and PhenoRAG-FT dependent on whether recall or precision is prioritized in a given clinical context. A detailed error analysis comparison of PhenoRAG and PhenoRAG-FT is provided in Appendix F.1

Table 5: Comparison of PhenoRAG and PhenoRAG-FT and benchmarks on the synthetic UTI dataset. Since the RAG-HPO models are non-deterministic (*), the reported values represent the average metrics across three runs.

Model	**Micro Average [%]**			**Macro Average [%]**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
PhenoBERT	65.79	35.21	45.87	51.94	34.67	41.58
PhenoBERT+CAN-BERT	75.76	35.21	48.08	55.83	34.67	42.77
RAG-HPO*	89.62	52.58	66.26	73.33	51.00	60.15
PhenoRAG	79.55	**98.59**	88.05	77.67	**95.00**	85.46
PhenoRAG+FT	**98.28**	80.28	**88.37**	**95.00**	82.72	**88.44**

6. Discussion

While detailed findings are presented in the results sections, we summarize here the key insights, limitations, and future directions.

PhenoRAG combines retrieval-augmented generation with a lightweight language model to deliver high-recall phenotype detection (crucial for clinical decision support) while re-

maining interpretable and suitable for local deployment. By isolating clinically relevant segments, it enables SLMs to match or surpass larger systems. Across real and synthetic datasets, PhenoRAG outperforms strong baselines, especially in recall, while maintaining solid precision and low computational cost. It effectively handles complex clinical language—negations, temporal cues, and nuanced symptom descriptions—where encoder-based models often struggle. Fine-tuning further improves specificity. Retrieval proved robust in our settings and enhances transparency by linking predictions to interpretable evidence.

Limitations PhenoRAG's current focus on predefined symptoms, consistent with guideline-based decision support systems, limits its applicability to large-scale or exploratory HPO mapping. While the hierarchical nature of the HPO supports partial generalization (e.g., capturing "paroxysmal lethargy" when querying for "lethargy"), this does not replace full concept coverage. Second, the lightweight nature of the "Medical Expert" SLM (chosen for local deployment) can occasionally confuse closely related symptoms (e.g., ureterolithiasis vs. flank pain). We mitigated this with fine-tuning (PhenoRAG-FT), though this adds training effort and computational cost. Third, PhenoRAG relies on successful retrieval and failure to retrieve relevant text prevents accurate SLM classification. In our evaluation, such cases were rare, but they remain a potential vulnerability.

Lastly, our real-world datasets, though diverse in setting and population, are relatively small. This limits robustness, especially for rare or unobserved phenotypes. Synthetic evaluations help but might not fully capture the complexity of clinical narratives and could introduce bias.

Outlook Future work could evaluate hyperparameters more systematically, particularly the retrieval cut-off ($S = 5$), which was selected based on prior work and computational efficiency. While this cut-off proved effective in our setting, other tasks may benefit from tuning this value or adopting adaptive retrieval strategies, especially as the number of target symptoms grows. Replacing synthetic sentences in the contextual database with real clinical text may also improve retrieval accuracy and reduce the number of segments needed. Additionally, considering larger datasets would enable more fine-grained analysis, improve generalizability, and guide model refinements. Finally, PhenoRAG could potentially be extended toward ontology-wide phenotype recognition by retrieving candidate terms per segment and using multi-label prompting. Although more computationally demanding, such extensions would increase flexibility and move the framework toward a general-purpose clinical tool.

7. Conclusion

PhenoRAG introduces an efficient, interpretable framework for phenotype identification, combining retrieval and generation to enhance contextual understanding using lightweight components. By narrowing the focus to relevant clinical extracts and deferring interpretation to a "Medical Expert" SLM, the method achieves strong performance (without training) with low resource demands. Validated on two real-world tasks—rare disease diagnosis and distinguishing asymptomatic from symptomatic bacteriuria—PhenoRAG shows strong clinical potential, particularly for deployment in resource-constrained settings.

Acknowledgments

AA and DC received funding from the grant #2021-911 of the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology). This work was supported by the Swiss National Science Foundation (grant no. 211422) to SDB. DSF is supported by the Swiss National Science Foundation [320030_231175 and 32030E_219127] and the University Research Priority program of the University of Zurich ITINERARE – Innovative Therapies in Rare Disease. The authors would like to thank Kathrin Zotter (University Children's Hospital Zurich) for her help in establishing the ground-truth phenotypes for the RD dataset.

References

Rune Aabenhus, Malene Plejdrup Hansen, Volkert Siersma, and Lars Bjerrum. Clinical indications for antibiotic use in danish general practice: results from a nationwide electronic prescription database. *Scandinavian journal of primary health care*, 35(2):162–169, 2017.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Abdulkadir Albayrak, Yao Xiao, Piyush Mukherjee, Sarah S Barnett, Cherisse A Marcou, and Steven N Hart. Enhancing human phenotype ontology term extraction through synthetic case reports and embedding-based retrieval: A novel approach for improved biomedical data annotation. *Journal of Pathology Informatics*, 16:100409, 2025.

Anthropic. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. Accessed: Apr 8th, 2025.

Anthropic. Claude 3.7 sonnet and Claude code. https://www.anthropic.com/news/claude-3-7-sonnet, 2025. Accessed: Apr 8th, 2025.

Aryan Arbabi, David R Adams, Sanja Fidler, Michael Brudno, et al. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR medical informatics*, 7(2):e12596, 2019.

Nicolas W Cortes-Penfield, Barbara W Trautner, and Robin Jump. Urinary tract infection and asymptomatic bacteriuria in older adults. *Infectious disease clinics of North America*, 31(4):673, 2017.

Yuhao Feng, Lei Qi, and Weidong Tian. Phenobert: a combined deep learning method for automated recognition of human phenotype ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1269–1277, 2022.

Brandon T Garcia, Lauren Westerfield, Priya Yelemali, Nikhita Gogate, E Andres Rivera-Munoz, Haowei Du, Moez Dawood, Angad Jolly, James R Lupski, and Jennifer E Posey. Improving automated deep phenotyping through large language models using retrieval augmented generation. *medRxiv*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tudor Groza, Harry Caufield, Dylan Gration, Gareth Baynam, Melissa A Haendel, Peter N Robinson, Christopher J Mungall, and Justin T Reese. An evaluation of GPT models for phenotype concept recognition. *BMC Medical Informatics and Decision Making*, 24(1): 30, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Martina Huemer, Daria Diodato, Bernd Schwahn, Manuel Schiff, Anabela Bandeira, Jean-Francois Benoist, Alberto Burlina, Roberto Cerone, Maria L Couce, Angeles Garcia-Cazorla, et al. Guidelines for diagnosis and management of the cobalamin-related remethylation disorders cblC, cblD, cblE, cblF, cblG, cblJ and MTHFR deficiency. *Journal of inherited metabolic disease*, 40:21–48, 2017.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, et al. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1): D1207–D1217, 2021.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Manuel Lobo, Andre Lamurias, and Francisco M Couto. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017 (1):8565739, 2017.

Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, and Zhiyong Lu. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, 37(13):1884–1890, 2021.

Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The lancet*, 399(10325):629–655, 2022.

Lindsay E Nicolle, Suzanne Bradley, Richard Colgan, James C Rice, Anthony Schaeffer, and Thomas M Hooton. Infectious diseases society of america guidelines for the diagnosis and treatment of asymptomatic bacteriuria in adults. *Clinical infectious diseases*, pages 643–654, 2005.

Jim O'Neill. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. *Rev. Antimicrob. Resist.*, 2014.

Adam C Palmer and Roy Kishony. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nature Reviews Genetics*, 14(4):243–248, 2013.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2020.

Md Abdus Salam, Md Yusuf Al-Amin, Moushumi Tabassoom Salam, Jogendra Singh Pawar, Naseem Akhter, Ali A Rabaan, and Mohammed AA Alqumber. Antimicrobial resistance: a growing serious threat for global public health. In *Healthcare*, volume 11, page 1946. Multidisciplinary Digital Publishing Institute, 2023.

Guillermo V Sanchez, Ahmed Babiker, Ronald N Master, Tony Luu, Anisha Mathur, and Jose Bordon. Antibiotic resistance among urinary isolates from female outpatients in the united states in 2003 and 2012. *Antimicrobial agents and chemotherapy*, 60(5):2680–2683, 2016.

Jean-Marie Saudubray, Matthias R Baumgartner, Ángeles García-Cazorla, and John H Walter. *Inborn metabolic diseases*. Springer, 2022.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

Gabrielle Stinton, Jane A. Lieviant, Sylvia Kam, Jiin Ying Lim, Jasmine Chew-Yin Goh, Weng Khong Lim, Gareth Baynam, Tele Tan, Duc-Son Pham, and Saumya Shekhar Jamuar. Clinical free text to hpo codes. *Rare*, 1:100007, 2023. ISSN 2950-0087. doi: 10.1016/j.rare.2023.100007. URL http://dx.doi.org/10.1016/j.rare.2023.100007.

Yvonne Su, Yonatan B Babore, and Charles E Kahn Jr. A large language model to detect negated expressions in radiology reports. *Journal of Imaging Informatics in Medicine*, pages 1–7, 2024.

Sanjeev K Swami, Juliette T Liesinger, Nilay Shah, Larry M Baddour, and Ritu Banerjee. Incidence of antibiotic-resistant escherichia coli bacteriuria according to age and location of onset: a population-based study from olmsted county, minnesota. In *Mayo Clinic Proceedings*, volume 87, pages 753–759. Elsevier, 2012.

Pugazhenthan Thangaraju and Sajitha Venkatesan. Who ten threats to global health in 2019: Antimicrobial resistance. *Cukurova Medical Journal*, 44(3):1150–1151, 2019.

Morgan R Timm, Seongmi K Russell, and Scott J Hultgren. Urinary tract infections: pathogenesis, host susceptibility and emerging therapeutics. *Nature Reviews Microbiology*, pages 1–15, 2024.

Betty Van Aken, Ivana Trajanovska, Amy Siu, Manuel Mayrdorfer, Klemens Budde, and Alexander Loeser. Assertion detection in clinical notes: Medical language models to the rescue? In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 35–40, 2021.

Andy Wang, Cong Liu, Jingye Yang, and Chunhua Weng. Fine-tuning large language models for rare disease concept normalization. *Journal of the American Medical Informatics Association*, 31(9):2076–2083, June 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae133. URL http://dx.doi.org/10.1093/jamia/ocae133.

Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns*, 5(1), 2024.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899, 2021.

Appendix A. Method

A.1. Contextual database generation

To identify the most relevant report segments for phenotype assessment by the "Medical Expert" SLM, each clinical report is first divided into smaller, manageable chunks and encoded using a sentence transformer. These encoded segments are then compared—via cosine similarity—to a set of synthetically generated, contextually enriched sentences representing each target phenotype. These synthetic sentences are produced using the same sentence transformer and serve as a reference for retrieval. The generation of these context-rich sentences follows the prompt proposed by Albayrak et al. (2025), shown in Figure 3.

> *"Generate <XX> unique sentences that describe the provided HPO label as they would appear in a clinical narrative or interpretive report. Each sentence should offer a different perspective or detail, similar to how various clinicians might report observations or diagnoses in clinical notes. Avoid using the exact phrase in every sentence to ensure diversity and reflect the range of clinical expression. Sometimes, I will provide comments, synonyms, and definitions that can help provide more context for your thoughts. Return a bulleted list instead of numbered. Feel free to use clinical shorthand that a physician might use in writing reports.*
>
> *HPO label: <Label>*
> *Definition: <Definition>*
> *Comments: <Comments>*
> *Synonyms: <Synonyms>"*

Figure 3: Prompt template used for generating contextual sentences for each target phenotype p_i to facilitate retrieval of relevant clinical report segments (adapted from (Albayrak et al., 2025)). Expressions enclosed in angle brackets (e.g., <Label>) are replaced with the corresponding phenotype-specific information from the HPO (Label, Definition, Comments, Synonyms) as well as the desired number of generated sentences (<XX>).

A.2. PhenoRAG-FT

In this paper we also introduce a fine-tuned variant of PhenoRAG—called PhenoRAG-FT—designed to further enhance the SLM's performance in particularly challenging or task-specific scenarios. Although the standard PhenoRAG performs well without additional fine-tuning, this optional low-resource fine-tuning approach can be used to address specific dataset complexities, such as distinguishing highly similar phenotypes or correctly interpreting clinical shorthand notations. PhenoRAG-FT achieves this by first generating a targeted synthetic dataset tailored explicitly to the difficulties at hand, and then fine-tuning the "Medical Expert" SLM specifically on this targeted dataset to enhance its contextual sensitivity and precision in phenotype identification. In this work, we suggest to target a better differentiation of closely related phenotypes—avoiding the inadvertent identification

of neighbouring symptoms—and enhanced robustness to typographical errors and complex negations. Nevertheless, this fine-tuning strategy can be easily adapted to other specialized tasks by adjusting the synthetic dataset generation accordingly.

The synthetic dataset creation is detailed below and illustrated in Figure 4.

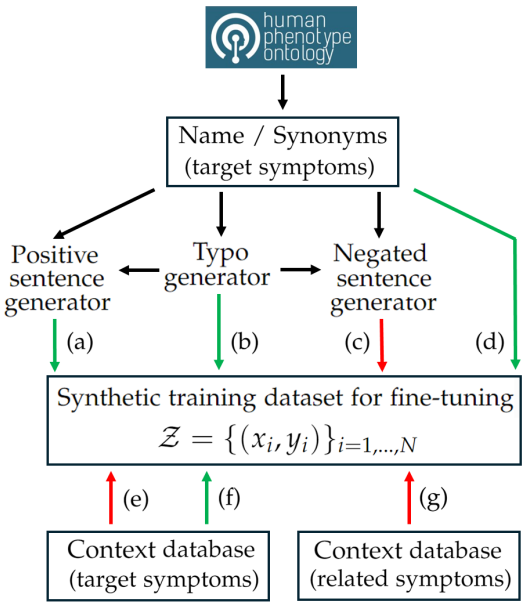

Figure 4: Overview of the synthetic data generation procedures (a-g) used to construct the training dataset for fine-tuning the "Medical Expert" SLM as part of PhenoRAG-FT. The arrows indicate the flow of information, while their colours represent the ingestion of positive (green) and negative (red) examples into the training dataset.

Dataset Initialization and Positive Examples: The synthetic dataset is initialized using sentences from the context database. This process yields a set of positive examples, each paired with a "Yes" ground-truth label.

Negative Example Generation: Negative examples are generated by intentionally mismatching phenotypes. For instance, a sentence corresponding to phenotype p_j is inserted into the prompt for symptom s_i (with $j \neq i$), and these examples are labeled as negative. Since these may represent only simple negatives—given the potential distinctiveness among target symptoms—the HPO ontology is further leveraged to generate hard negative examples based on neighbouring concepts. Using the same procedure as for the initial context database, a set of sentences is generated for each closely related term with a decoder-only generative language model and the context sentence generation prompt (see Figure 3).

Augmenting with Typographical Errors and Negations: To enhance practical relevance, the dataset is augmented with examples that include both typographical errors and negated statements. Typographical errors are introduced via a "Typo Generator" function, which randomly flips letters in a word or swaps them with neighboring ones on the keyboard, as proposed by Wang et al. (2024). By applying this function to symptom names and their synonyms, additional positive examples with intentional spelling mistakes are pro-

duced. Negated examples are generated using a set of pre-defined template sentences that explicitly denote the absence of a symptom. The "Negated Sentence Generator" function embeds the target symptom into a randomly chosen negation template—either using the original term or one altered by the typo generator. Similarly, a "Positive Sentence Generator" augments the dataset with additional confirming sentences, again with and without spelling errors. Finally, the direct use of symptom names and synonyms as keyword-style positive examples further reinforces the model's exposure to the shorthand notation frequently encountered in clinical practice.

Appendix B. Benchmarks

B.1. Preliminary Benchmark Evaluation

To determine a suitable baseline, we conducted a preliminary evaluation of existing phenotype extraction methods on the GSC+ dataset (Lobo et al., 2017). As shown in Figure 5, rule-based (MetaMap) and CNN-based (NeuralCR) approaches underperformed compared to transformer-based models. While PhenoTagger slightly outperformed PhenoBERT by up to 0.9 percentage points across metrics, the performance gap was minimal. Given PhenoBERT's competitive accuracy and simplicity of integration, we selected it as our primary baseline for further evaluation.

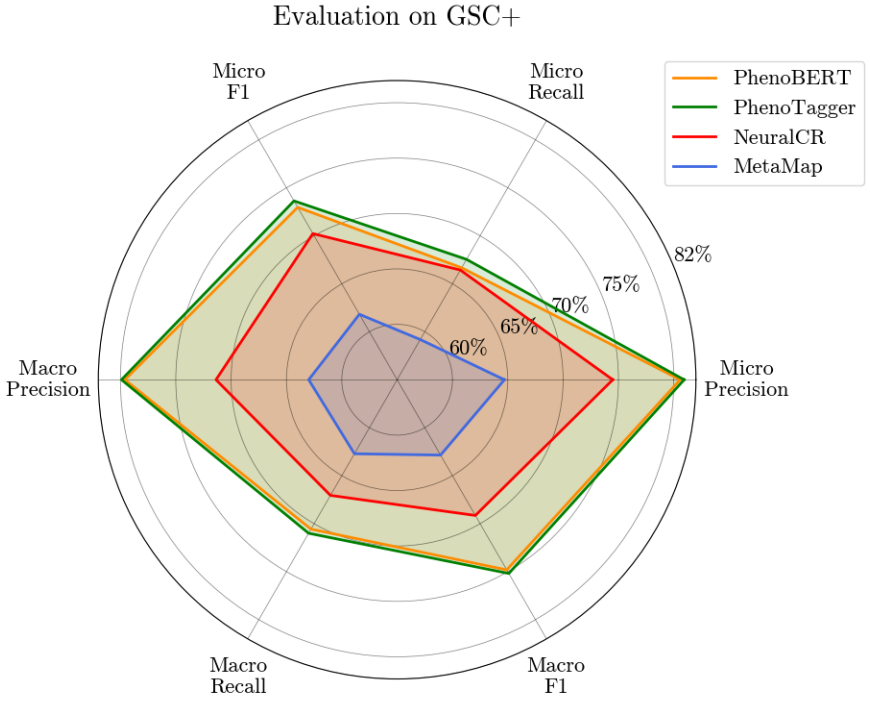

Figure 5: Performance comparison of rule-, CNN-, and BERT-based models. The exact values underlying this visualization are provided in Table 6).

Table 6: Performance comparison of rule-, CNN-, and BERT-based models on GSC+.

Model	Micro Average [%]			Macro Average [%]		
	P	R	F1	P	R	F1
PhenoBERT	80.57	66.67	72.96	79.60	70.57	74.81
PhenoTagger	**80.93**	**67.54**	**73.63**	**79.91**	**71.00**	**75.19**
NeuralCR	74.49	66.43	70.23	71.37	67.05	69.14
MetaMap	64.70	59.19	61.82	62.99	62.71	62.85

The rule-based MetaMap reached the lowest values across all metrics. NeuralCR achieved a micro F1-score of 70.2%, surpassing MetaMap by 8.4 percentage points. This improvement was driven by a performance gap of 7.2 to 9.8 points in precision and recall. Similarly, NeuralCR also outperformed MetaMap on the macro level by 6.3 percentage points in F1. However, it falls short of the performance of both PhenoTagger and PhenoBERT across all metrics. While PhenoTagger achieved the highest values in all categories, the differences to PhenoBERT range from 0.3 to 0.9 percentage points.

B.2. PhenoBERT+CAN-BERT

As preliminary experiments revealed that PhenoBERT particularly struggles with negation detection, we enhanced it by integrating the Clinical Assertion and Negation Classification BERT (CAN-BERT) model (Van Aken et al., 2021), resulting in the PhenoBERT+CAN-BERT variant. CAN-BERT, a transformer-based model fine-tuned on the i2b2 corpus, classifies entities as "positive", "possible", or "absent" based on contextual cues. It has been shown to outperform rule-based systems like medSpaCy in identifying negations and improving clinical information extraction (Su et al., 2024). Given its proven effectiveness and compatibility with clinical texts, its integration into our pipeline offers a more competitive benchmark for evaluating phenotype detection (see Figure 6).

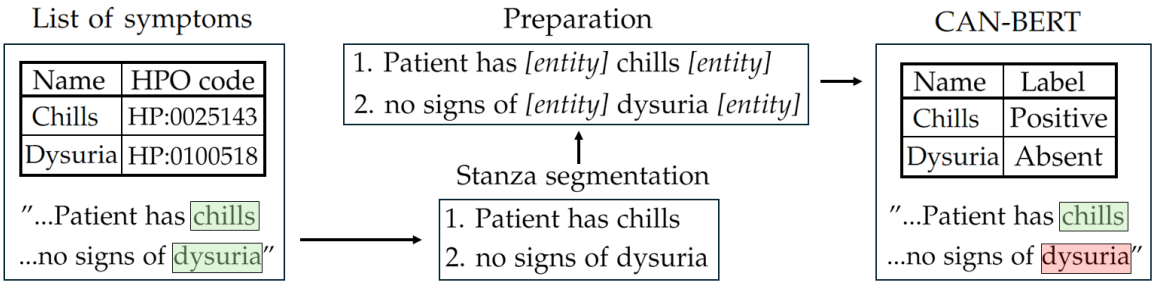

Figure 6: Integration of the negation detection with CAN-BERT. This figure serves as an example illustrating the process. The colouring indicates the respective meta-annotation. Entries marked in green are regarded as 'Positive' while the concept highlighted in red is viewed as 'Absent' by the model. The initial list of symptoms and the corresponding meta-annotations are obtained with PhenoBERT.

CAN-BERT was accessed from Hugging Face on 30/11/2024. Since the model is restricted to an input size of 512 tokens, the reports must be segmented before passing them to the model. For this step it is essential to preserve a maximum of clinically relevant context within each fraction of the text in order to facilitate the downstream classification. Therefore, the segmentation was performed with the *mimic*-package of Stanza, which is a dedicated deep learning-based NLP tool trained on clinical data (Qi et al., 2020; Zhang et al., 2021). Based on this preprocessing step, the integration of CAN-BERT then leverages the information provided by the initial model. Per default, the annotation pipeline outputs a list of concepts along with the positional information of their trigger words. This information is used to insert the token "[entity]" before and after the trigger words to mark the corresponding concept. After enclosing each concept with the designated tokens, the corresponding sentences are forwarded to the model for processing. CAN-BERT then provides the meta-annotation for each concept based on the contextual information contained in the relevant text segment. Only concepts labeled as "positive" are included in the final set of annotations.

Appendix C. Evaluation

Model performance is evaluated using macro and micro precision, recall, and F1-score, based on the predicted concepts $\hat{Y}_i$ and ground truth Y_i for each report i among N total reports.

Macro-level metrics are computed by averaging per-report scores, treating all reports equally:

$$\text{Macro Precision} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \frac{|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i|} \tag{1}$$

$$\text{Macro Recall} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \frac{|\hat{Y}_i \cap Y_i|}{|Y_i|} \tag{2}$$

$$\text{Macro F1} = \frac{2 \times \text{Macro Precision} \times \text{Macro Recall}}{\text{Macro Precision} + \text{Macro Recall}} \tag{3}$$

Micro-level metrics aggregate predictions across all reports, weighting by the number of annotations:

$$\text{Micro Precision} = \frac{\sum_{i=1}^{\mathrm{N}} |\hat{Y}_i \cap Y_i|}{\sum_{i=1}^{\mathrm{N}} |\hat{Y}_i|} \tag{4}$$

$$\text{Micro Recall} = \frac{\sum_{i=1}^{\mathrm{N}} |\hat{Y}_i \cap Y_i|}{\sum_{i=1}^{\mathrm{N}} |Y_i|} \tag{5}$$

$$\text{Micro F1} = \frac{2 \times \text{Micro Precision} \times \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}} \tag{6}$$

If the ground truth does not contain any HPO concepts for a specific report, the calculation of Micro Recall results in a division by zero. In these cases, the Micro Recall was set

to one if the model also produced an empty set of annotations. Otherwise, the Micro Recall was defined as zero. Similarly, if the model did not extract any concepts while the ground truth was non-empty, the Micro Precision was set to zero. In cases where both the model and the ground truth contained empty sets of annotations, the Micro Precision was set to one. These conventions align with previous work Feng et al. (2022); Luo et al. (2021).

Appendix D. Datasets

D.1. Remethylation Disorder Dataset

Table 7 provides a summarized overview of the remethylation disorder cohort, whose clinical reports were used in the rare disease experiments.

Table 7: Remethylation Disorder Cohort Characteristics.

Biochemical Presentation	**N**	**Diagnosis**	**Count**
elevated Hcy	44	Severe *MTHFR* deficiency	22
		Unknown	22
elevated Hcy and MMA	74	cbl*C* deficiency (c.394C>T)	26
		cbl*C* deficiency (c.271dupA)	26
		Unknown	22

Table 8 lists the occurrence counts of the target phenotypes in the rare disease (RD) dataset, while Table 9 provides the description of each phenotype. Seizures and developmental delay were the most frequently reported phenotypes, each appearing in 40 of the clinical reports.

Table 8: Target phenotypes for the RD dataset and their occurrences in clinical reports (ordered by decreasing count)

Phenotype	**HPO Code**	**Count**
Seizure	HP:0001250	40
Neurodevelopmental delay	HP:0012758	40
Hypotonia	HP:0001252	28
Anemia	HP:0001903	28
Feeding difficulties	HP:0011968	18
Failure to thrive	HP:0001508	17
Cerebral atrophy	HP:0002059	15
Hypertonia	HP:0001276	15
Thrombocytopenia	HP:0001873	13

Table 9: Target Phenotypes for the RD dataset and Their Occurrences in Clinical Reports

Phenotype	HPO Code	Description
Feeding difficulties	HP:0011968	Impaired ability to eat related to problems gathering food and getting ready to suck, chew, or swallow it.
Failure to thrive	HP:0001508	Child whose physical growth is substantially below the norm.
Seizure	HP:0001250	Intermittent abnormality of nervous system physiology characterized by a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain.
Hypotonia	HP:0001252	Abnormally low muscle tone.
Hypertonia	HP:0001276	Increased muscle tone.
Neurodevelopmental delay	HP:0012758	Delay in one or more developmental domains compared to typical development.
Anemia	HP:0001903	A reduction in erythrocytes volume or hemoglobin concentration.
Thrombocytopenia	HP:0001873	Low platelet count.
Cerebral atrophy	HP:0002059	Atrophy (wasting, decrease in size of cells or tissue) affecting the cerebrum.

D.2. Hospital UTI dataset

The pipeline used for phenotype-based UTI classification is illustrated in Figure 7. The UTI-specific phenotypes that are targeted in this experiment are detailed in Table 10.

D.3. Synthetic UTI Dataset

The synthetic UTI dataset for further validating PhenoRAG relied on the prompt in Figure 8. Some statistics about the generated synthetic reports can be found in Tables 11 and 12.

Appendix E. Additional Results

E.1. Rare Disease Experiment

Table 13 presents specific examples where PhenoRAG's predictions differ from those of the baseline models. These examples support the error analysis by illustrating the distinct strengths and limitations of each approach. In particular, they highlight the contextual understanding capabilities of PhenoRAG, which help explain its substantially higher recall compared to PhenoBERT.

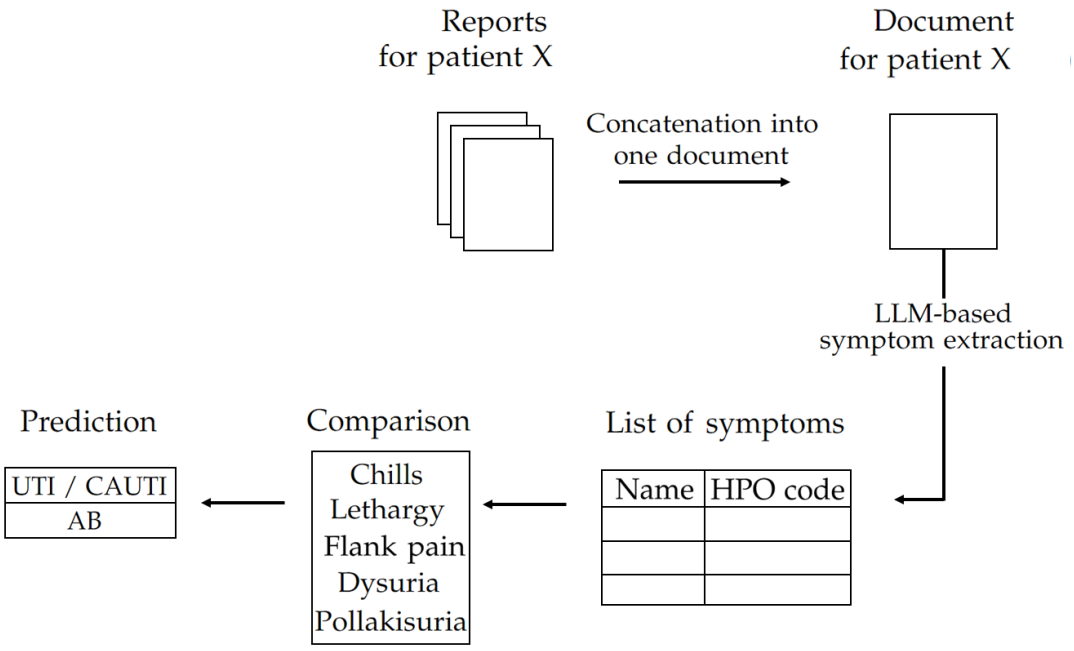

Figure 7: Visualization of the pipeline for symptom-based report classification.

Appendix F. UTI Detection Experiment

F.1. Comparison or PhenoRAG and PhenoRAG-FT

Fine-tuning, which incorporates hard negatives, typographical errors, and complex negations, refines the model's mapping of text to symptom definitions. For example, the default PhenoRAG model occasionally mislabels related conditions—such as mapping "ureterolithiasis" to "flank pain"—and overgeneralizes descriptions, whereas PhenoRAG-FT enforces stricter boundaries, correctly distinguishing between closely associated symptoms like costovertebral angle tenderness and flank pain, or differentiating nocturia from pollakisuria. Although this precision improvement sometimes reduces recall for certain symptoms (e.g., lethargy and flank pain), likely due to limited fine-tuning data, the overall reduction of silent errors enhances clinical reliability. Ultimately, the choice between the raw and fine-tuned models depends on the specific use case, with high-supervision settings favouring precision and unsupervised scenarios benefiting from the broader recall of the default model.

Table F in Appendix E provides specific examples of sentences where the fine-tuning made a difference. 1) *Related symptoms* resemble the target but correspond to different symptoms. For example "increased tendency to wake up to urinate" which is nocturia but resembles "pollakisuria" which is the symptom picked up by PhenoRAG while PhenoRAG-FT correctly captures the difference.

2) *Causal symptoms* may contribute to the target but represent separate conditions. For instance, nephritis usually causes flank pain but doesn't mean that the patient has flank pain. While PhenoRAG interpreted the segment "confirmed the presence of nephritis" as flank pain, this was correctly discarded by its fine-tuned variant.

Table 10: Target Phenotypes for UTI Identification

Phenotype	HPO Code	Synonyms	Description
Dysuria	HP:0100518	Painful or difficult urination; Dull burning sensation with urination	Painful or difficult urination.
Pollakisuria	HP:0100515	Constant urination; Frequent urination	Increased frequency of urination.
Flank pain	HP:0030157	Kidney pain	An unpleasant sensation characterized by physical discomfort (such as pricking, throbbing, or aching) and perceived to originate in the flank.
Chills	HP:0025143	/	A sudden sensation of feeling cold.
Lethargy	HP:0001254	Dullness; Inaction; Inactivity; Languor; Lethargy; Slowness; Torpor	A state of fatigue, either physical or mental slowness and sluggishness, with difficulties in initiating or performing simple tasks. Distinguished from apathy which implies indifference and a lack of desire or interest in the task. A person with lethargy may have the desire, but not the energy to engage in personal or socially relevant tasks.

Table 11: Synthetic UTI Dataset: Report-Level Statistics

Metric	Value
Report Length (chars)	Mean: 2259 Min: 809 Max: 6147 STD: 1291

Table 12: Synthetic UTI Dataset: Phenotype Occurrence Summary

Phenotype	Occurrences
Lethargy	18
Pollakisuria	17
Flank Pain	14
Dysuria	11
Chills	11

3) *Joint negations* Through targeted improvement of complex negations, PhenoRAG-FT was able to correctly discard the symptom lethargy from the segment "no evidence of pollakisuria, chills, or lethargy" which was otherwise missed by the out-of-the-box PhenoRAG.

However the fine-tuning also sometimes deteriorated the performance of PhenoRAG. Correct symptom descriptions, *nagging pain on the right side, between ribs and hips* and *non-specific fatigue, routine activities feel effortful despite motivation* which were correctly captured as Flank Pain and Lethargy, respectively by PhenoRAG, were discarded after the fine-tuning procedure.

"Generate a comprehensive clinical report describing the state of a patient. Below, I provide details about five HPO symptoms. Embed information about the presence or absence of these symptoms into the report. The embedded information can be subtle and may include abbreviations, complex negations or paraphrases of the symptom. Be creative in making it challenging to detect the symptoms. Afterward, state which symptoms you chose to be present in the report.

HPO label: Chills
Definition: A sudden sensation of feeling cold.
Comments: The word chills can also refer to an episode of shivering, accompanied by paleness and feeling cold.

HPO label: Lethargy
Definition: A state of fatigue, either physical or mental slowness and sluggishness, with difficulties in initiating or performing simple tasks. Distinguished from apathy which implies indifference and a lack of desire or interest in the task. A person with lethargy may have the desire, but not the energy to engage in personal or socially relevant tasks.
Comments: Apathy and lethargy may co-occur.
Synonyms: Dullness, Inaction, Inactivity, Languor, Lethargy, Slowness, Torpor

HPO label: Flank pain
Definition: An unpleasant sensation characterized by physical discomfort (such as pricking, throbbing, or aching) and perceived to originate in the flank.
Comments: The flank is the area on the side of the abdomen between the ribs and the hip.
Synonyms: Kidney pain

HPO label: Dysuria
Definition: Painful or difficult urination.
Synonyms: Dull burning sensation with urination, Painful or difficult urination

HPO label: Pollakisuria
Definition: Increased frequency of urination.
Synonyms: Constant urination, Frequent urination

Figure 8: Complete prompt template for generating synthetic clinical reports for the synthetic UTI dataset.

Table 13: Relevant text segments from the RD dataset, ground truth phenotype, and model predictions. A checkmark (✓) indicates a correct prediction. Incorrect predictions are reported with the prediction made by the model. "None" means that no phenotype was identified in the segment. "NDD" stands for Neurodevelopmental Delay. For PhenoBERT, we ignore HPO codes that differ from target phenotypes.

Text Segment	**Ground Truth**	**Predictions**		
		PB	**PB+CB**	**PhenoRAG**
no evidence of megaloblastic anemia	None	Anemia	✓	✓
mum reports a developmental delay initially but this has picked up in the last two months	None	NDD	NDD	✓
can sit with support only, never walked or acquired speech	NDD	None	None	✓
height-related body weight of 9.2 kg was below the normal range (<P3)	Failure to thrive	None	None	✓
Brain MRI findings: [...], mild generalized atrophy	Cerebral atrophy	None	None	✓
poor feeding at the age of ten days	Feeding difficulties	None	None	✓
appearing somewhat muscle-hypotonic	Hypotonia	None	None	✓
could not be fed orally	Feeding difficulties	None	None	✓
One was a male infant with psychomotor retardation [...].	None (sibling)	NDD	NDD	NDD
Intra-uterine growth delay.	None	✓	✓	Failure to thrive

Table 14: Text segments illustrating the issues that can lead to false-positive predictions by the PhenoBERT models. PB stands for the default implementation while PB+CB denotes the enhancement with CAN-BERT.

Text segment from reports	**PB**	**PB+CB**	**Issue**
no evidence of lethargy	Lethargy	None	*Simple* negation
patient confirms that chills have not occurred	Chills		*Complex* negation
return if symptoms develop, including chills, ...	Chills		Potential future state
inability to urinate	Pollakisuria		Erroneous mapping
Additional examples			
his father frequently had flank pain due to kidney stones	Flank pain		Family history
dysuria has disappeared	Dysuria		Past state

Table 15: Relevant text segments from patient reports and the corresponding symptoms identified by the models. PB refers to the original PhenoBERT model, while PB+CB denotes its enhancement with CAN-BERT.

Text segment	**PB**	**PB+CB**	**PhenoRAG**	**PhenoRAG-FT**
increased urinary frequency	None		Pollakisuria	
localized pain in the left flank	Pain		Flank pain	
burning sensation during micturition	None		Dysuria	
occasional sensations of coldness	None		Chills	

Table 16: Text segments illustrating the distinct prediction characteristics of the default RAG implementation (PhenoRAG) and the fine-tuned model (PhenoRAG-FT). The first column contains exemplary text segments while the second column lists the corresponding symptom. The last two columns represent the answers from the models. The ground truth answer for each row is highlighted in bold italic font.

Related symptoms	**Target**	**PhenoRAG**	**PhenoRAG-FT**
tenderness upon deep palpation of the costovertebral angle	Flank pain	Present	***Absent***
increased tendency to wake up to urinate	Pollakisuria	Present	***Absent***
Causal symptoms			
confirmed the presence of nephritis	Flank pain	Present	***Absent***
urinary bladder inflammation with evidence of bacterial infection	Dysuria	Present	***Absent***
Joint negations			
no evidence of pollakisuria, chills, or lethargy	Lethargy	Present	***Absent***
no fever, chills, or weight loss	Chills	Present	***Absent***
Correct symptom descriptions			
nagging pain on the right side, between ribs and hips	Flank pain	***Present***	Absent
non-specific fatigue, routine activities feel effortful despite motivation	Lethargy	***Present***	Absent