A Pipeline for Automated Phenotype Extraction from Medical Reports using Large Language Models

Andrea Bombarda[1] [a], Martina Saletta[2] [b], Matteo Bellini[3] [c], Lucrezia Goisis[3] [d], Maria Iascone[3] [e], Paolo Cazzaniga[2] [f], Domenico Fabio Savo[1] [g]

[1]*Department of Management, Information, and Production Engineering, University of Bergamo, Bergamo, Italy*

[2]*Department of Human and Social Sciences, University of Bergamo, Bergamo, Italy*

[3]*Medical Genetics Lab, ASST Papa Giovanni XXIII, Bergamo, Italy*

{*andrea.bombarda, domenicofabio.savo, martina.saletta, paolo.cazzaniga*}*@unibg.it*, {*m.bellini, miascone*}*@asst-pg23.it*, *lucreziagoisis@hotmail.it*

Keywords: Large Language Models, Medical Genetics, Phenotype Extraction, Rare Genetic Disease Research

Abstract: Unstructured clinical narratives are a major source of phenotypic evidence for rare-disease diagnosis and genomic variant interpretation. However, their free-text nature, often multilingual, heterogeneous in format, and inconsistent in terminology, makes automated phenotype extraction and interoperability with downstream genomic pipelines difficult. This creates a practical bottleneck for scalable and reproducible phenotype curation in medical genetics, where manual review is time-consuming and prone to variability. To address this problem, we propose a robust, open-source, and fully local pipeline for automatically extracting and standardizing patient phenotypes from medical reports while preserving data privacy. The pipeline integrates: (i) OCR-based digitization and an LLM-based translation module to produce an English version of the report; (ii) a GPT-oss–based phenotype extractor using structured, few-shot prompting to identify phenotypes relevant to the index patient; and (iii) a fuzzy standardization stage that combines lexical similarity with embedding-based semantic matching to map extracted phenotypes to Human Phenotype Ontology (HPO) concepts. Our multi-stage design improves robustness to real-world documentation issues, including multilingual acronyms, variable report structure, spelling errors, and synonym variability, and it ensures privacy compliance by keeping all computation on local infrastructure. We demonstrate the pipeline end-to-end on a representative clinical report, showing that it extracts patient-relevant phenotypes and produces HPO-aligned, machine-readable outputs suitable for downstream genomic analyses. This work provides a practical foundation for privacy-preserving, scalable phenotype curation in clinical genetics and supports future integration and evaluation on larger clinical datasets.

1 INTRODUCTION

Clinical reports written in natural language contain a wealth of valuable information that can assist medical professionals in diagnosing conditions, identifying patterns, and making informed decisions (Maguluri, 2024; Zhou et al., 2025). These reports encapsulate detailed descriptions of patient symptoms, observations, and diagnostic results, reflecting the complexity and diversity of human diseases in the real-world. Despite their value, the unstructured nature of clinical narratives presents significant challenges for automated analysis, preventing large-scale computational studies and integration with structured biomedical data (Sedlakova et al., 2023). For example, clinical reports may contain the description of symptoms of multiple subjects (e.g., the patients and their family) or refer to the same pathology with different synonyms (e.g., headache or hemicrania).

In this paper, we target the identification of phenotypes in clinical reports, i.e., the set of observable characteristics or traits of an organism (Wojczynski and Tiwari, 2008). This work has been developed in the context of a national project, ANTHEM (Fon-

[a] https://orcid.org/0000-0003-4244-9319
[b] https://orcid.org/0000-0001-7250-8987
[c] https://orcid.org/0009-0001-4297-9160
[d] https://orcid.org/0009-0007-2665-8096
[e] https://orcid.org/0000-0002-4707-212X
[f] https://orcid.org/0000-0001-7780-0434
[g] https://orcid.org/0000-0002-8391-8049

dazione ANTHEM, 2022), aimed at enhancing and introducing advanced technologies within the processes followed by physicians in the medical genetics department of the Papa Giovanni XXIII Hospital, in Bergamo (Italy)[1]. Automated extraction and standardization of phenotypic information from clinical reports is a critical task to realize and implement data-driven medicine. On the one hand, manual curation, while accurate, is resource-intensive, time-consuming, and infeasible on the scale of modern clinical datasets, which routinely encompass thousands of patient records (Kang et al., 2025). On the other hand, unstructured data slows research and clinical decision-making, prevents the identification of disease patterns, and hinders the consistent and automatic aggregation of data.

In a genetic laboratory, the extraction of phenotypes from medical reports is a critical enabler for efficient and reproducible variant interpretation. After DNA sequencing, bioinformatics processing identifies possible variations that may cause genetic diseases. The geneticists then must determine whether any of these variants plausibly explain the patient's condition, which clinically manifests itself through the observed phenotypes documented in the clinical report. Automating the identification and normalization of these phenotypes reduces the reliance on time-consuming manual curation and ensures consistent use of the phenotype evidence in the downstream analyses (Yuan et al., 2022; Pandey et al., 2012; Deisseroth et al., 2019). By structuring phenotypic features early in the workflow, laboratories can improve their outcome and accelerate reporting timelines while maintaining traceability and interoperability with existing genomic pipelines.

Recent advances in Natural Language Processing (NLP) technology have dramatically reshaped the landscape of biomedical data analysis. Large Language Models (LLMs), such as ChatGPT and LLaMa, have shown an exceptional ability to interpret natural language, opening new possibilities for extracting complex structured information from textual data (Saletta et al., 2025). Additionally, a great effort has been put in by practitioners into designing a set of ontologies. This is the case of the Human Phenotype Ontology (HPO) (Köhler et al., 2020), which catalogs all known phenotypes and their relations. The use of such an ontology is crucial for mapping extracted clinical concepts to universally defined terms, facilitating interoperability, reliability, and reproducibility in clinical and genetic research. It also defines unique identifiers for each phenotype (including possible synonyms), which are used in third datasets to eliminate ambiguities when referring to specific phenotypes.

This work presents a comprehensive pipeline designed to address the challenge of automatically extracting phenotypic information from clinical reports. This pipeline leverages LLMs and fuzzy matching techniques to map phenotypes to HPO terms.

The remainder of this paper is structured as follows. Section 2 provides the necessary background on the genetics process and on the importance of identifying phenotypes in clinical texts for genetics research. Section 3 presents our pipeline, while Section 4 provides a demonstration of its functioning, reports the results of a preliminary evaluation, and discusses possible limitations. Finally, Section 5 and Section 6 discuss related works and conclude the paper, respectively.

2 BACKGROUND

This section presents an in-depth overview of the genetic process, emphasizing the stages in which the proposed pipeline we introduce in this paper can be employed to enhance it.

A medical genetics laboratory typically operates through two tightly connected processes: generating genomic data by sequencing patient DNA and interpreting identified genetic variants to explain the patient's clinical phenotype and potential genetic pathologies.

Initially, a genetics laboratory receives a clinical request with detailed phenotype information written in natural language (and, possibly, in any language) in the form of a clinical report, collects an appropriate specimen (often blood or saliva), and isolates high-quality genomic DNA under standardized conditions. In the context in which we operate, genetics laboratories frequently investigate potential pathologies in children. They consider not only the DNA and clinical information of the index case (i.e., the child) but also those of parents and relatives.

On one side, the DNA is processed through a sequencing workflow designed to detect genomic variants. On the other side, genetics specialists review the clinical report to identify the relevant phenotypes it contains. We define *relevant* phenotypes as those pertaining to the index case and potentially informative for explaining the observed clinical condition.

Then, scientists and medical geneticists evaluate each variant in the context of the patient's phenotype and family history. Variants are annotated (Bombarda

[1]The Papa Giovanni XXIII Hospital ranked 117^{th} in the World's Best Hospital ranking in 2025: https://rankings.newsweek.com/worlds-best-hospitals-2025

et al., 2025) with information from public and curated databases (such as population frequencies, disease associations, and functional annotations) and relevant literature, and are then classified as pathogenic, likely pathogenic, likely benign, or benign.

Identifying the *relevant* phenotypes in a clinical report is crucial for the work of a genetics laboratory because phenotypic information provides the clinical context needed to guide and prioritize genetic analyses. Accurate phenotypes help laboratory professionals select the most appropriate testing strategies, interpret variants in light of the patient's observable traits, and distinguish potential pathogenic findings from benign ones. They also support effective genotype–phenotype correlation, enabling more precise diagnostic conclusions and improving communication with clinicians. Phenotype identification, despite its significance, is inherently subjective and prone to errors. Different geneticists may have varying opinions on what constitutes a relevant phenotype, and some phenotypes may be overlooked altogether.

To address this problem, in this paper, we propose a robust, open-source, and locally hosted pipeline to extract and standardize phenotypes from medical reports w.r.t. the HPO ontology.

2.1 HUMAN PHENOTYPE ONTOLOGY

The Human Phenotype Ontology (HPO) (Köhler et al., 2020) is a structured, machine-readable vocabulary designed to represent human phenotypic abnormalities in a standardized and hierarchically organized manner. Each HPO term describes a specific clinical feature and is associated with a unique and stable identifier. Examples of terms, with their own identifier, are microcephaly (HP:0000252[2]), ataxia (HP:0001251[3]), or epilepsy (HP:0001250[4]). Terms are linked through well-defined relationships (e.g., is-a or part-of), which enable the representation of phenotypic knowledge at varying levels of granularity. In addition to the ontology structure itself, HPO provides curated annotations linking phenotypic terms to genetic disorders, genes, and disease models, forming one of the most widely used resources for computational phenotype analysis in rare disease diagnostics and research.

Adopting such an ontology is essential in clinical genomics for several reasons. First, it enables consistent and unambiguous communication of phenotypic information across clinicians, laboratories, and bioinformatics tools, reducing the variability inherent in free-text clinical descriptions. Second, its interoperability with other biomedical ontologies and databases enhances integration within clinical pipelines, supporting reproducible workflows and enabling large-scale analyses of phenotypic and genomic data.

[2]https://hpo.jax.org/browse/term/HP:0000252

[3]https://hpo.jax.org/browse/term/HP:0001251

[4]https://hpo.jax.org/browse/term/HP:0001250

3 OUR PIPELINE

In this section, we present our pipeline for extracting HPO-standardized phenotypes from clinical reports. Figure 1 depicts such a pipeline. We designed our solution to comply with all GDPR requirements (Feretzakis et al., 2025): All data is processed locally, without using any cloud or remote service.

Designing a local LLM-based pipeline for phenotype extraction from medical reports is essential because it ensures that sensitive patient data remains securely within the institution's infrastructure. Unlike cloud-based solutions, a local pipeline avoids data-sharing risks and complies more easily with strict privacy regulations governing clinical records. Moreover, running the entire workflow on-premise allows the genetics laboratory to tailor prompts, fine-tune models, and adapt extraction logic to its specific reporting style, medical domain, and diagnostic needs.

Our pipeline exploits GPT-oss, with 120B parameters (Bi et al., 2025), running over Ollama[5]. We initially tested other open-source and locally hosted LLMs, e.g., LLaMa 3.2 (Grattafiori et al., 2024), LLaMa 3.3, LLaMa 4, Gemma 3 (Team et al., 2025), and DeepSeek v3 (DeepSeek-AI et al., 2025). However, during an exploratory and preliminary study, we found that GPT-oss was the LLM leading to the best precision and recall. All interactions with our LLM are made using a null temperature to ensure that only original information is used, and no additional phenotypes are added by the model. We tested our pipeline on a PC running Ubuntu Server 24.04.2, with 256GB RAM, AMD Ryzen Threadripper PRO 7985WX 64-Cores CPU, and a GeForce RTX 4090 GPU.

In the following, we describe in detail each step of the pipeline reported in Figure 1, namely the translation phase (Section 3.1), the extraction of relevant phenotypes (Section 3.2), and their standardization (Section 3.3).

Replication material and source code are available at https://github.com/ANTARES-PRJ/DataExtraction.

[5]https://ollama.com/

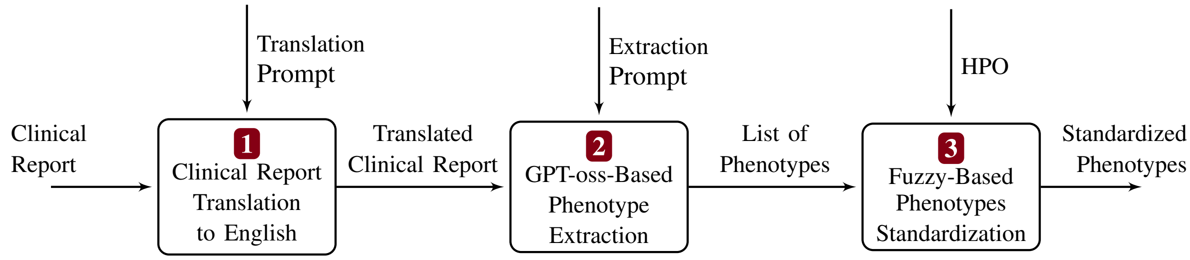

Figure 1: Our phenotype extraction pipeline.

3.1 TRANSLATION

In medical settings, clinical reports are often written in the language of the country in which the hospitals operate. For this reason, we introduce the translation step (1 in Figure 1) to translate the content of the clinical report into English. While this step is optional, it is recommended due to the significantly higher performance of LLMs when working with English text compared to other languages (Qin et al., 2025; Zhang et al., 2024).

This phase is composed of two main tasks:

- **OCR-based text extraction**: The textual content is extracted from the clinical report, which is typically provided as a scanned PDF or image. To accomplish this, we employ the `py-tesseract` library.
- **Text translation**: The extracted text is then translated from its original language into English.

Regarding the translation step, our initial approach considered the use of standard translation services, such as the Google Translate APIs. Nevertheless, two potential drawbacks emerged. First, relying on online services such as the Google APIs would prevent the pipeline from processing all data locally, thereby introducing external data handling and potential privacy concerns. It may also result in service unavailability when the Internet connection fails, preventing the system from meeting the required service level. Second, we observed that medical reports may contain acronyms that differ across languages. For instance, Amyotrophic Lateral Sclerosis is abbreviated as ALS in English, whereas in French, Italian, and Spanish it is written as SLA. Using a translator that understands the context—such as an LLM (Huang et al., 2023; Shu et al., 2024)—is, thus, paramountly important.

To address these issues, we implement in this phase a GPT-oss-based translator. It receives the digitalized content of a clinical report as input and, using the *translation prompt* reported in Figure 2, it generates a *translated clinical report*. It is worth mentioning that the pipeline was tested considering only medical reports written in Italian. It is therefore necessary to assess the quality of the translation in the case where texts written in other languages are used, especially less-common or low-resource can be adapted to adapt to other languages ones. In any case, the translator integrated into our pipeline can be substituted or the *translation prompt* can be adapted to better fit to other languages, if required.

```
Translate the following clinical report into
English.
Do not summarize, interpret, expand, modify,
or infer any information.
Preserve all medical terminology, formatting,
and numerical values exactly as they appear,
translating only the natural-language text.
Respond with the full translation and nothing
else.

Clinical report:
{Text}
```

Figure 2: The translation prompt.

3.2 PHENOTYPE EXTRACTION

Once the clinical report has been translated into English, we carry out the phenotype extraction step (2 in Figure 1). This phase is crucial, as the accuracy of all subsequent processes depends on the correctness and completeness of the extracted phenotypes. In particular, this step must address several challenges:

- **Heterogeneous report formats**: Clinical reports do not follow a uniform structure. Different hospitals may include custom headers or footers that contain references to phenotypes unrelated to the specific patient under examination.
- **Multiple patients described in a single document**: A clinical report may include phenotype descriptions for more than one individual. For instance, it is common to report the phenotypes of the index case and those of the parents together.

- **Terminological variability**: The same phenotype can be expressed using different terms or lexical variants (e.g., *epileptic seizure*, *seizures*, and *epilepsy* may all refer to the same clinical concept). Additionally, typos may be present in human-written text.

In theory, phenotypes could be identified through straightforward substring matching against a dictionary of phenotype names. However, due to the three challenges outlined above, this approach is insufficient, and a context-aware extraction method is required instead. Thus, we here propose to extract phenotypes by using an LLM-based approach as we do with step 2 in our pipeline.

Our phenotype extractor is fed with the *translated clinical report* produced by the activity 1 and with the *extraction prompt* reported in Figure 3. Our prompt provides a detailed description of the task, organizes it into multiple sub-tasks, and provides an example of a medical report from which to extract the desired output. This approach, known as few-shot prompting (Bahrami et al., 2023), is commonly employed to enhance the performance of LLMs.

As a result, this phase outputs a list of *relevant* phenotypes in JSON format, where each entry includes the fields `Original_phenotype_name` and `HPO_phenotype_name`. The former corresponds to the exact wording used in the clinical report, while the latter denotes the phenotype name that GPT-oss associates with an existing term in the HPO ontology. Note that we include both fields in our pipeline output as our preliminary tests demonstrated that, due to errors or hallucinations (Das et al., 2025), LLMs sometimes provide phenotype names that do not exist in the HPO ontology.

3.3 PHENOTYPE STANDARDIZATION

After extracting the *relevant* phenotypes from the clinical report, we apply a standardization step (3 in Figure 1) to align these phenotypes with the HPO ontology and retrieve their corresponding unique identifiers. Indeed, despite the ability of GPT-oss in identifying relevant phenotypes, we observed that the model sometimes returns incorrect or incomplete HPO phenotype names. To mitigate this issue, our pipeline includes an additional step, which standardizes the extracted phenotypes via fuzzy matching. This step searches for the closest match, when available, of each term extracted from the report in HPO, considering both preferred phenotype names and their synonyms. This process combines lexical and semantic similarity measures to obtain robust matches even in the presence of spelling variations, synonyms, or differing phrase structures.

First, we generate semantic embeddings for all HPO terms and synonyms using the `all-MiniLM-L6-v2` model. These embeddings provide a dense vector representation of each phenotype, enabling similarity-based retrieval in a continuous vector space. We emphasize that the semantic embeddings are stored and generated only when the pipeline is used the first time. All the other interactions with the pipeline do not require these embeddings to be generated.

For every phenotype extracted by phase 2, we compute two complementary similarity scores:

- **Fuzzy string similarity**: We apply the `fuzz.token_sort_ratio` method from the `fuzzywuzzy` library to measure lexical similarity between the extracted phenotype and each HPO term. This metric yields an integer value between 0 and 100, where 100 denotes a perfect lexical match. We call this score s_{fuzzy}.
- **Cosine similarity on embeddings**: We encode the extracted phenotype using the same embedding model and compute its cosine similarity with all HPO embeddings. This captures semantic proximity even when the phrasing differs significantly. We call this score s_{cosine}.

The final similarity score, s_{final} is obtained by combining the two metrics:

$$s_{final} = 0.5 \cdot \left(\frac{s_{fuzzy}}{100}\right) + 0.5 \cdot s_{cosine}$$

Using the final similarity score, we standardize the extracted phenotype terms. To accommodate near-matches while filtering out overly vague ones, we define two distinct similarity thresholds. First, we introduce a lower bound for similarities, θ_l. All matches below this threshold are automatically discarded, as they are too different from all HPO entries. Similarly, we introduce an upper bound θ_u. All matches above this threshold are considered perfect matches. In our case, we set $\theta_l = 0.8$ and $\theta_u = 0.9999$. We empirically defined these two values. First, $\theta_l = 0.8$ allows for providing high confidence in the output provided by our pipeline. Given that s_{final} considers both string similarity and embeddings similarity, setting θ_l to a relatively high value allows for retaining only values that are strongly similar overall, not just weakly related. Instead, $\theta_u = 0.9999$ considers perfect matches by avoiding floating-point edge cases while approximating 1.0 (Goldberg, 1991; Rump et al., 2008).

Given the similarity score and the two bounds, three outcomes are possible:

```
Please carefully read the following medical report about a patient. Your task involves multiple
steps to process and extract specific information related to the patient's phenotypes as detailed
 below:

1. Extract Patient Phenotypes: Identify and extract only the phenotypes described for the patient
. Do not include information related to relatives or other individuals. Focus solely on the
patient's conditions and characteristics as mentioned in the report.

2. Map to HPO Terms: For each identified phenotype, find the corresponding Human Phenotype
Ontology (HPO) term. If a direct match for an HPO term cannot be found, identify the closest
related term.

3. Structured Output: Present your findings in JSON format: for each identified phenotype, you
should provide:
    - Original_phenotype_name: The name or description of the phenotype as identified by you in
    the patient's medical report;
    - HPO_phenotype_name: The matched or closest HPO term;

Your output must only include the results in JSON format with the fields specified above, without
 any additional commentary or information. Hence, for instance, in a medical report such as the
following "The patient shows epileptic behavior during the day, after having lunch. In the past,
she registered a rise in blood pressure at the 37th gestational week. Her mother was reported to
have facial edema" the output must be as follows:
[{
    "Original_phenotype_name" : "Seizure",
    "HPO_phenotype_name": "Seizure"
},{
    "Original_phenotype_name" : "Rise in blood pressure at 37th gestational week",
    "HPO_phenotype_name": "Hypertension"
}]

No comment is allowed after the quotes, neither with parenthesis or double slashes. The text is:
{Text}
```

Figure 3: The extraction prompt.

- **No match**: No HPO term has a score s_{final} exceeding a predefined similarity threshold θ_l, leading to no standardized result provided by our pipeline.
- **Single match**: Exactly one HPO term has a score s_{final} surpassing the threshold θ_l; this term is selected as the standardized phenotype, and it is reported as an exact match if $s_{final} > \theta_u$.
- **Multiple matches**: More than one HPO term achieves a sufficiently high score $s_{final} > \theta_l$; in this case, all candidates are returned and expert validation is required to determine the most appropriate mapping.

As for phase 2, the final output of this phase is a JSON list in which each entry contains not only the fields `Original_phenotype_name` and `HPO_phenotype_name`, but also the associated similarity `score`, the matched `HPO_term`, the `HPO_matched_phenotype` (which could be a synonym and not the real phenotype name), and the corresponding `HPO_URI`.

4 EVALUATION

In this section, we provide a demonstration of how our pipeline works, starting from a clinical report to the final list of HPO-standardized phenotypes, we report the results of a preliminary evaluation, and discuss possible limitations. To avoid including non-English text, we omit the translation phase (1) in this demonstration and, instead, rely on the already translated clinical report shown in Figure 4. This clinical report describes three different phenotypes, i.e., *Headache* (HP:0002315), *nocturia* (HP:0000017), and *seizure* (HP:0001250). Note that this report contains additional phenotypes, which are not related to the patient. The patient's mother is reported to have *ear pain* (HP:0030766), and the report comes from a department specialized in *sleep disturbances* (HP:0002360) and *insomnia* (HP:0100785).

Once the translated report is available, the phenotype extraction phase (2) can start. To perform this phase, the prompt in Figure 3 is completed by appending the translated report in Figure 4. We report the re-

St. Aurelia General Hospital – City of Veloria
Date: November 15, 2025
Clinical Report

The patient reports recurrent episodes of intense headache over the past three months, often accompanied by brief generalized seizures. Nocturia has also been noted, with the patient waking multiple times during the night to urinate
.
His mother reports having frequent ear pain.
Further diagnostic evaluation is recommended.

Dr. Mickey Mouse
Department of Sleep Disorders and Insomnia

Figure 4: The translated clinical report.


```
[
  {
    "Original_phenotype_name": "Intense
    ↪ headache",
    "HPO_phenotype_name": "Headache"
  },
  {
    "Original_phenotype_name": "Brief
    ↪ generalized seizures",
    "HPO_phenotype_name": "Seizure"
  },
  {
    "Original_phenotype_name": "Nocturia",
    "HPO_phenotype_name": "Nocturia"
  }
]
```


Figure 5: The list of phenotypes extracted from the clinical report.

sults of this phase in Figure 5. It can be observed that only the phenotypes pertaining to the specific patient are extracted. This shows that our pipeline is effective in filtering non-relevant phenotypes. For each of them, the `Original_phenotype_name` field reflects the wording used in the clinical report, while the `HPO_phenotype_name` field corresponds to the term that the LLM identifies as the appropriate HPO entry.

The list of phenotypes is then passed to the standardization phase (3), which considers HPO and maps each term in Figure 5 to HPO concepts. The result of this phase is reported in Figure 6. For each phenotype identified in the previous phase, this step enriches the output with additional information, including the details of the corresponding HPO entry, the URI that links to further phenotype-specific documentation (e.g., associated diseases), and the computed similarity score s_{final}.

Although simple, this example illustrates the overall functioning of the pipeline and shows its effectiveness in the extraction of structured information from


```
[
 {
  "Original_phenotype_name": "Intense
  ↪ headache",
  "HPO_phenotype_name": "Headache",
  "score": 1.0,
  "HPO_matched_term": "Headache",
  "HPO_matched_phenotype": "Headache",
  "HPO_matched_URI":
  ↪ "http://purl.obolibrary.org/obo
     /HP_0002315"
 },
 {
  "Original_phenotype_name": "Brief
  ↪ generalized seizures",
  "HPO_phenotype_name": "Seizure",
  "score": 1.0,
  "HPO_matched_term": "Seizure",
  "HPO_matched_phenotype": "Seizure",
  "HPO_matched_URI":
  ↪ "http://purl.obolibrary.org/obo
     /HP_0001250"
 },
 {
  "Original_phenotype_name": "Nocturia",
  "HPO_phenotype_name": "Nocturia",
  "score": 1.0,
  "HPO_matched_term": "Nocturia",
  "HPO_matched_phenotype": "Nocturia",
  "HPO_matched_URI":
  ↪ "http://purl.obolibrary.org/obo
     /HP_0000017"
 }
]
```


Figure 6: The list of standardized phenotypes.

unstructured medical reports. In the following, we report the results of a preliminary evaluation on a more extensive dataset.

Preliminary evaluation To provide a preliminary assessment of our pipeline's effectiveness, we compared it with PhenoGPT (Yang et al., 2024), an advanced phenotype recognition approach based on the LLaMa 2 model, with 7B parameters. We conducted this evaluation using the BioLarkGSC+ dataset (Lobo et al., 2017), which contains 228 clinical reports along with their corresponding phenotype annotations. Our preliminary results show that, compared with PhenoGPT, our pipeline achieves a 6% improvement in precision, a 24% improvement in recall, and a 12% increase in F1-score.

Limitations Despite the promising results, the proposed pipeline has several limitations that should be considered when interpreting its outputs and when deploying it in operational genetics workflows. Our pipeline assumes that the clinical report text can be reliably extracted via OCR from scanned PDFs or images. OCR errors (e.g., due to poor scan quality, complex layouts, or unusual fonts) may propagate to downstream steps, ultimately affecting phenotype extraction and normalization. In addition, the pipeline relies on a translation-to-English stage (recommended because LLMs perform better on English text) which, as discussed, may introduce mistranslations that can impact phenotype recognition. While we tested our system in Italian, and it turned out to work properly, we cannot exclude these issues during in real-world use.

While our pipeline uses a null temperature to reduce the risk of "invented" content, the possibility of extraction mistakes cannot be fully eliminated in a fully automated setting. Similarly, we cannot rule out the possibility of extracting non-patient-related phenotypes: disentangling which phenotypes belong to which individual may fail when narrative structure is ambiguous or when multiple individuals are discussed extensively. We envision our system as a support for genetics researchers and not to fully replace them, at least in the preliminary steps, where further evaluation may be needed. Also, to enable fully autonomous use of the proposed pipeline in real-world clinical settings, the software would need to undergo certification. (Bombarda et al., 2022; Bombarda et al., 2021; Waldock et al., 2024; De Vito, 2024)

The standardization step (Section 3.3) allows multiple matches, in which more than one HPO term exceeds the similarity threshold. In such cases, the pipeline returns all candidates and requires expert validation to select the correct mapping. Therefore, the pipeline is not fully autonomous in difficult cases and still depends on domain expertise for disambiguation.

Finally, while our experience demonstrates that the pipeline can be hosted locally, it still requires substantial RAM and a high-end GPU. This may limit adoption in settings without comparable hardware resources or require lighter model variants that could reduce extraction quality.

5 RELATED WORK

The automatic extraction of phenotypic information from clinical narratives has been extensively studied within biomedical informatics, particularly with the rise of electronic health records and the need for large-scale computational analyses.

Early approaches relied on rule-based systems and dictionary matching, which provided limited flexibility in handling linguistic variability and context, such as MetaMap (Aronson and Lang, 2010) and cTAKES (Savova et al., 2010). The former maps free text to UMLS Metathesaurus concepts and has been widely used for extracting clinical phenotypes. The latter is an open-source pipeline that uses dictionary lookup and context-dependent rules to recognize clinical entities and normalize them to UMLS codes. However, these tools were not specifically targeting phenotypes and struggled with ambiguous phrasing, multilingual data, and heterogeneous document structure—limitations that directly impact information extraction. MedCAT (Kraljevic et al., 2019) adopted a similar approach by providing flexible dictionary-based extraction for any ontology and supports user feedback to improve accuracy. Nevertheless, while achieving high precision, MedCAT may miss context nuances.

In the domain of phenotype-specific extraction, several methods have focused on structured ontologies, such as the HPO, the same ontology that we use in our pipeline. ClinPhen (Deisseroth et al., 2019) leverages HPO to map symptoms from clinical notes to standardized terms. However, this tool relies on deterministic matching and classifier-based techniques, which makes it sensitive to phrasing, synonyms, typographical errors, and language variations. Similarly, HPO is used by SapBERT (Abdulnazar et al., 2023), where a pretrained BERT is further trained by pulling synonyms (e.g., from HPO) closer in embedding space, yielding a model that achieves state-of-the-art concept normalization accuracy.

Recent advances in Natural Language Processing have introduced transformer-based architectures and proposed the use of Large Language Models (LLMs) for clinical information extraction. Clinical adaptations of BERT, such as BioBERT (Lee et al., 2019), ClinicalBERT (Huang et al., 2020), and Bio+ClinicalBERT (Ling, 2023), substantially improved performance in named entity recognition, relation extraction, and phenotype extraction tasks. These LLMs are used in many tools, such as Phenotagger (Luo et al., 2021) or PhenoBCBERT (Yang et al., 2024). Similarly, GPT-like structures have been used in PhenoGPT (Yang et al., 2024) and BioGPT (Al-Kateb et al., 2025). However, while these tools demonstrated good performance in information extraction, to the best of our knowledge, they have been tested only on English inputs, and they do not provide any standardized output, as we do with our approach.

More recently, general-purpose LLMs, such as GPT-3.5, GPT-4, and Mistral have shown strong capabilities in interpreting long and complex clinical narratives (Singhal et al., 2023; Bhattarai et al., 2024). However, despite their high accuracy, many LLM-based approaches rely on cloud-based APIs, raising data privacy concerns that limit applicability in genetics laboratories and other clinical workflows (Chen et al., 2025). Unlike these approaches, we use a locally-hosted and open-source LLM to avoid any privacy-related issue.

Our effort aims at extracting and standardizing phenotypes from unstructured clinical text. This activity can be leveraged by new diagnostic AI models for differential diagnoses of rare diseases, as proposed by PhenoBrain (Mao et al., 2025).

6 CONCLUSION

This work presents a fully local, reproducible, and privacy-preserving pipeline for automated phenotype extraction from unstructured clinical reports. By combining OCR-based preprocessing, Large Language Model–driven information extraction, and a robust fuzzy-based standardization mechanism grounded in the Human Phenotype Ontology (HPO), the proposed approach addresses several long-standing challenges in clinical text processing, including heterogeneous document structures, multilingual content, terminological variability, and inconsistent phenotype naming. Unlike traditional rule-based or dictionary-driven systems, our pipeline leverages the contextual reasoning capabilities of a large language model to accurately identify phenotypes relevant to the index patient, even in the presence of complex narrative structures or references to multiple individuals. At the same time, the use of a locally hosted and open-source LLM ensures full compliance with privacy and data protection requirements, a critical constraint in real-world medical genetics laboratories that limits the applicability of many existing cloud-based solutions. The subsequent standardization phase further enhances the reliability and usability of the extracted information by aligning free-text phenotype descriptions with HPO concepts through a combination of lexical and semantic similarity measures. This design choice mitigates common sources of error such as spelling variations, synonym usage, and partial matches, and produces structured and interoperable outputs suitable for downstream genomic analysis and phenotype-driven variant interpretation.

Through a concrete demonstration, we showed how the pipeline successfully extracts and standardizes clinically relevant phenotypes while filtering out non-patient-related information, highlighting its practical applicability in realistic clinical scenarios. Overall, this work contributes a modular and extensible solution that bridges the gap between unstructured clinical narratives and structured phenotype representations, supporting more efficient, consistent, and scalable phenotype curation in medical genetics workflows.

Future work will focus on conducting an extensive quantitative evaluation against manually annotated clinical datasets, the integration of domain-specific fine-tuning or RAG (Celsa et al., 2025) strategies to further improve extraction accuracy, and the extension of the pipeline to support larger corpora and additional clinical ontologies, such as Mondo (Vasilevsky et al., 2022). We will also evaluate the performance of the pipeline on medical reports written in different languages, in order to assess the effectiveness of the translation component and to potentially provide users with the option to select the translator that best suits their specific needs. Finally, we plan to investigate human-in-the-loop validation strategies to assist clinicians in resolving ambiguous phenotype mappings and to further increase trust and adoption in clinical practice.

Acknowledgments

This work was funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them. We acknowledge Samuele Negri for conducting a preliminary evaluation of this pipeline as part of his master's thesis.

REFERENCES

Abdulnazar, A., Kreuzthaler, M., Roller, R., and Schulz, S. (2023). *SapBERT-Based Medical Concept Normalization Using SNOMED CT*. IOS Press.

Al-Kateb, G., Cengiz, E., and Gök, M. (2025). Biogpt: A generative transformer-based framework for personalized genomic medicine and rare disease diagnosis. *Mesopotamian Jour-*

nal of Artificial Intelligence in Healthcare, 2025:154–164.

Aronson, A. R. and Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Bahrami, M., Mansoorizadeh, M., and Khotanlou, H. (2023). Few-shot learning with prompting methods. In *2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 1–5.

Bhattarai, K., Oh, I. Y., Sierra, J. M., et al. (2024). Leveraging gpt-4 for identifying cancer phenotypes in electronic health records: a performance comparison between gpt-4, gpt-3.5-turbo, flant5, llama-3-8b, and spacy's rule-based and machine learning-based methods. *JAMIA Open*, 7(3):ooae060.

Bi, Z., Chen, K., Tseng, C.-Y., et al. (2025). Is gpt-oss good? a comprehensive evaluation of openai's latest open source models.

Bombarda, A., Bellini, M., Iascone, M., and Savo, D. (2025). A flexible and open-source tool for genetic variant annotation. In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies*, page 406–413. SCITEPRESS - Science and Technology Publications.

Bombarda, A., Bonfanti, S., Galbiati, C., Gargantini, A., Pelliccione, P., Riccobene, E., and Wada, M. (2021). Lessons learned from the development of a mechanical ventilator for covid-19. In *2021 IEEE 32nd International Symposium on Software Reliability Engineering (ISSRE)*, pages 24–35.

Bombarda, A., Bonfanti, S., Galbiati, C., Gargantini, A., Pelliccione, P., Riccobene, E., and Wada, M. (2022). Guidelines for the development of a critical software under emergency. *Information and Software Technology*, 152:107061.

Celsa, C., Giuffrè, M., Di Maria, G., et al. (2025). Pheno-rag: An artificial intelligence tool for guideline-informed management decision of hepatocellular carcinoma. *JHEP Reports*, page 101715.

Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., and Wu, P. (2025). A survey on privacy risks and protection in large language models. *Journal of King Saud University Computer and Information Sciences*, 37(7).

Das, A. B., Ahmed, S., and Sakib, S. K. (2025). Hallucinations and key information extraction in medical texts: A comprehensive assessment of open-source large language models.

De Vito, G. (2024). Assessing healthcare software built using iot and llm technologies. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, EASE '24, page 476–481, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Liu, A., Feng, B., et al. (2025). Deepseek-v3 technical report.

Deisseroth, C. A., Birgmeier, J., Bodle, E. E., et al. (2019). Clinphen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genetics in Medicine*, 21(7):1585–1593.

Feretzakis, G., Vagena, E., Kalodanis, K., Peristera, P., Kalles, D., and Anastasiou, A. (2025). Gdpr and large language models: Technical and legal obstacles. *Future Internet*, 17(4):151.

Fondazione ANTHEM (2022). ANTHEM web page. https://fondazioneanthem.it/en/.

Goldberg, D. (1991). What every computer scientist should know about floating-point arithmetic. *ACM Comput. Surv.*, 23(1):5–48.

Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). The llama 3 herd of models.

Huang, H., Wu, S., Liang, X., et al. (2023). Towards making the most of llm for translation quality estimation. In *Natural Language Processing and Chinese Computing*, pages 375–386, Cham. Springer Nature Switzerland.

Huang, K., Altosaar, J., and Ranganath, R. (2020). Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Kang, Y.-J., Lee, H., Yi, J. P., et al. (2025). Large language model versus manual review for clinical data curation in breast cancer: Retrospective comparative study. *JMIR Medical Informatics*, 13:e73605–e73605.

Kraljevic, Z., Bean, D., Mascio, A., et al. (2019). Medcat – medical concept annotation tool.

Köhler, S., Gargano, M., Matentzoglu, N., et al. (2020). The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ling, Y. (2023). Bio+clinical bert, bert base, and cnn performance comparison for predicting drug-review satisfaction.

Lobo, M., Lamurias, A., and Couto, F. M. (2017). Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017:1–8.

Luo, L., Yan, S., Lai, P.-T., et al. (2021). Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, 37(13):1884–1890.

Maguluri, K. K. (2024). *Natural language processing in healthcare: Unlocking insights from clinical data*. Deep Science Publishing.

Mao, X., Huang, Y., Jin, Y., et al. (2025). A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehrs. *npj Digital Medicine*, 8(1).

Pandey, K. R., Maden, N., Poudel, B., Pradhananga, S., and Sharma, A. K. (2012). The curation of genetic variants: Difficulties and possible solutions. *Genomics, Proteomics & Bioinformatics*, 10(6):317–325.

Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., and Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, 6(1):101118.

Rump, S. M., Ogita, T., and Oishi, S. (2008). Accurate floating-point summation part i: Faithful rounding. *SIAM Journal on Scientific Computing*, 31(1):189–224.

Saletta, M., Bombarda, A., Bellini, M., Goisis, L., Cazzaniga, P., Iascone, M., and Savo, D. F. (2025). *Automated Phenotype-Based Clustering of Clinical Reports Using Large Language Models*, page 345–350. Springer Nature Switzerland.

Savova, G. K., Masanz, J. J., Ogren, P. V., et al. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Sedlakova, J., Daniore, P., Horn Wintsch, A., et al. (2023). Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digital Health*, 2(10):e0000347.

Shu, P., Chen, J., Liu, Z., et al. (2024). Transcending language boundaries: Harnessing llms for low-resource language translation.

Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Team, G., Kamath, A., Ferret, J., et al. (2025). Gemma 3 technical report.

Vasilevsky, N. A., Matentzoglu, N. A., Toro, S., et al. (2022). Mondo: Unifying diseases for the world, by the world. *medRxiv*.

Waldock, W. J., Zhang, J., Guni, A., et al. (2024). The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: Systematic review and meta-analysis. *J Med Internet Res*, 26:e56532.

Wojczynski, M. K. and Tiwari, H. K. (2008). *Definition of Phenotype*, page 75–105. Elsevier.

Yang, J., Liu, C., Deng, W., Wu, D., Weng, C., Zhou, Y., and Wang, K. (2024). Enhancing phenotype recognition in clinical notes using large language models: Phenobcbert and phenogpt. *Patterns*, 5(1):100887.

Yuan, X., Wang, J., Dai, B., et al. (2022). Evaluation of phenotype-driven gene prioritization methods for mendelian diseases. *Briefings in Bioinformatics*, 23(2).

Zhang, Z., Zhao, J., Zhang, Q., Gui, T., and Huang, X. (2024). Unveiling linguistic regions in large language models. In *Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 6228–6247. Association for Computational Linguistics.

Zhou, S., Xu, Z., Zhang, M., et al. (2025). Large language models for disease diagnosis: a scoping review. *npj Artificial Intelligence*, 1(1).