Phenotypes Extraction from Text: Analysis and Perspective in the LLM Era

Moussa Baddour, Stéphane Paquelet, Paul Rollier, Marie De Tayrac, Olivier Dameron, Thomas Labbé

Proceedings of 2024 IEEE 12th International Conference on Intelligent Systems (IS)

Phenotypes Extraction from Text: Analysis and Perspective in the LLM Era

1st Moussa BADDOUR
Institute of Research and Technology b<>com
Rennes, France
Moussa.BADDOUR@b-com.com

2nd Stéphane PAQUELET
b<>com
Rennes, France
Stephane.PAQUELET@b-com.com

3rd Paul ROLLIER
University Hospital of Rennes
Rennes, France
Paul.ROLLIER@chu-rennes.fr

4th Marie DE TAYRAC
University Hospital of Rennes
Rennes, France
Marie.DE.TAYRAC@chu-rennes.fr

5th Olivier DAMERON
Univ Rennes, Inria, CNRS, IRISA - UMR 6074
Rennes, France
Olivier.DAMERON@univ-rennes.fr

6th Thomas LABBE
Orange
Rennes, France
Thomas.LABBE@orange.com

***Abstract*—Collecting the relevant list of patient phenotypes, known as deep phenotyping, can significantly improve the final diagnosis. As textual clinical reports are the richest source of phenotypes information, their automatic extraction is a critical task. The main challenges of this Information Extraction (IE) task are to identify precisely the text spans related to a phenotype and to link them unequivocally to referenced entities from a source such as the Human Phenotype Ontology (HPO).**

Recently, Language Models (LMs) have been the most successful approach for extracting phenotypes from clinical reports. Solutions such as PhenoBERT, relying on BERT or GPT, have shown promising results when applied to datasets built on the hypothesis that most phenotypes are explicitly mentioned in the text. However, this assumption is not always true in medical genetics. Hence, although the LMs carry powerful semantic abilities, their contributions are not clear compared to syntactic string-matching steps that are used within the current pipelines.

The goal of this study is to improve phenotype extraction from clinical notes related to genetic diseases. Our contributions are threefold: First, we provide a clear definition of the phenotype extraction task from free text, along with a high-level overview of the involved functions. Second, we conduct an in-depth analysis of PhenoBERT, one of the best existing solutions, to evaluate the proportion of phenotypes predicted with simple string-matching. Third, we demonstrate how utilizing and incorporating large language models (LLMs) for span detection step can improve performance especially with implicit phenotypes. In addition, this experiment revealed that the annotations of existing dataset are not exhaustive, and that LLM can identify relevant spans missed by human labelers.

***Index Terms*—phenotype,genetic, entity linking, phenoBERT, LLM, embeddings**

I. Introduction

Biomedical Entity Linking (BEL) [1] is a core natural Language Processing (NLP) task in the biomedical field. It acts as a bridge between the unstructured text and structured knowledge bases. It consists in finding and connecting biomedical concepts, terms, and entities mentioned in medical texts to their matching entries in structured databases or referenced ontologies.

Linking medical terms to the Human Phenotype Ontology (HPO[1]) [2] is an important asset for rare diseases diagnosis as it helps diagnose conditions more accurately, improves genetic testing, and accelerates researches.

BEL [1] task can be straightforward when the target entities are explicitly mentioned in the text with the same (or almost the same) labels as the ones in the knowledge base. It becomes more complex when the explicit mention does not match the target surface form (e.g. “lipid myopathy” in the text should be mapped to the referenced entity “Increased muscle lipid content”). It turns to be particularly challenging when the target is implicitly mentioned (e.g. the whole sentence “he is not independent when it comes to dressing and undressing” refers to the target entity “Intellectual disability”), which is often the case for the phenotypes in clinical reports related to rare genetic diseases. In the current paper, we present precisely the target task and the existing solutions, before analyzing more deeply PhenoBERT [5] which is an open-sourced solution reaching state-of-the-art performance. We then propose to modify one step of its pipeline with a LLM [4] approach in order to tackle some observed limitations. Finally, a qualitative analysis allows to highlight further insights regarding the dataset Ground Truth (GT).

II. Background

A. Task Description

Extracting specific information from unstructured data is a common task in NLP and information retrieval. However, the lack of predefined structure makes automated systems that seek to understand and extract logical concepts difficult to build. Meeting this challenge requires a systematic approach, which typically involves several key steps: span detection, candidate retrieval, and candidate ranking (Figure 1).

[1]http://www.human-phenotypeontology.org

979-8-3503-5098-2/24/$31.00 ©2024 IEEE

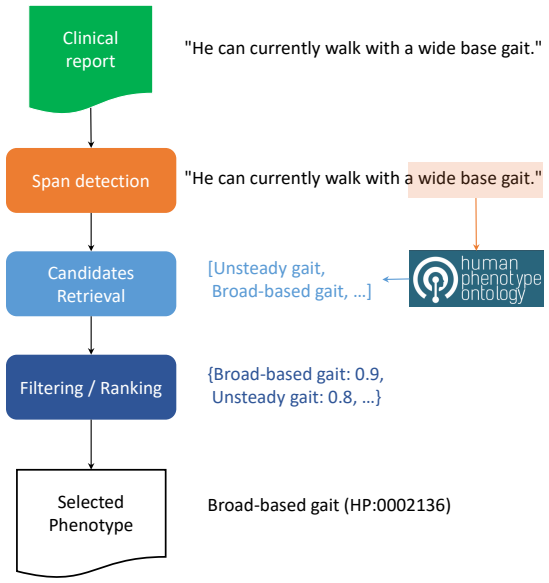

Fig. 1. High-level functional view of Biomedical Entity Linking.

1) Span detection: this first step involves finding specific text spans containing information of interest. These spans can be related to patients' conditions ("patient reported onset of chest pain two days ago", "history of hypertension"), observable symptoms ("fever", "nausea", "headache"), diagnosis elements ("elevated fasting blood glucose levels", "presence of glycosuria"), particular behaviors ("he hardly talks with others") or other relevant information. Some techniques such as Named Entity Recognition (NER) [3] are commonly used to detect and classify these spans based on linguistic patterns and contextual indicators.

2) Candidates Retrieval: once spans have been identified, the next step is to retrieve candidate entities that may match these spans. This process usually involves searching in databases, knowledge bases, or other information systems. Retrieval techniques can vary depending on the target knowledge (e.g. number of classes) and target application constraints (real-time vs offline), from simple keyword matching to more complex techniques such as semantic similarity with embeddings.

3) Candidates Ranking: The candidates are ranked by default based on the retrieval algorithm (e.g. Levenshtein distance, cosine similarity...). Most of the time, this ranking has to be improved with additional strategies involving filtering or re-ranking combining multiple criteria such as specificity, in-context relevance, in-domain frequency, etc., leading to more accurate confidence scores. There are several factors that contribute to ranking, including specificity, relevance, frequency, semantic similarity, and confidence scores. By integrating some of these factors, we ensure that the selected HPO term not only accurately represents the clinical feature but also aligns with the broader clinical context.

B. Existing solutions

Clinical concept recognition has progressed from rule-based methods that rely on predefined rules and knowledge bases to more advanced machine learning and deep learning approaches that use neural networks and word embeddings for enhanced pattern recognition. Recently, the introduction of transformer models, LLMs [4], and hybrid methods, which combine traditional techniques with deep learning, has further improved accuracy and efficiency.

Traditional clinical concept recognition tools relied on dictionary or rule-based methods like MetaMap [9], NCBO Annotator [10], OBO Annotator [11], ClinPhen [12], and Doc2HPO [13]. These tools employ various strategies, from knowledge-intensive mapping to sequential analytic procedures, to recognize phenotypes within biomedical texts. MetaMap, developed at the National Library of Medicine, is a program that connects biomedical texts to the Metathesaurus using a knowledge-intensive method or equivalently to discover Metathesaurus concepts referred to in the text. The NCBO (National Center for Biomedical Ontology) annotator initially provides direct annotations from raw text based on syntactic concept recognition, using terms from UMLS (Unified Medical Language System) and NCBO BioPortal ontologies. The Open Biological and Biomedical Ontologies (OBO) annotator is specifically implemented to annotate biomedical literature with HPO phenotypic abnormalities. It can also be applied to recognize terms from any OBO ontology, as it is mainly a named entity recognizer, which matches input text against terms from an OBO ontology ClinPhen is a tool designed to identify and extract clinically relevant phenotypic information from medical data. It utilizes a rule-based NLP system combined with sequential analytic procedures to distinguish accurate mentions of phenotypes from false positives. Doc2Hpo is an interactive web application that enables efficient phenotype concept curation from clinical text with automated concept normalization using the HPO.

Recently, researchers have increasingly favored the adoption of machine learning models, particularly deep learning architectures such as Convolutional Neural Networks (CNNs) [18] and Recurrent Neural Networks (RNNs) [19], due to their heightened accuracy and reduced reliance on hand-crafted features. The Neural Concept Recognizer (NCR) [20] is a tool designed to annotate unstructured text with concepts from an ontology. NCR employs a convolutional neural network trained using fastText [21] word vectors derived from the HPO ontology to encode input phrases. It then ranks medical concepts based on their similarity to the HPO ontology. Additionally, NCR can generalize to synonyms not explicitly included in the training data.

Transformers, illustrated by BERT (Bidirectional Encoder Representations from Transformers) [14], have garnered significant attention from researchers due to their parallelization capabilities and adeptness in recognizing long-range relationships. PhenoTagger [23] is a hybrid method that com-

bines dictionary and deep learning-based (BERT) methods to recognize HPO concepts in unstructured biomedical text. PhenoBERT [5] is a hybrid method that uses advanced deep learning methods (LSTMs [24] + Dictionary based + CNN + BERT) to identify clinical disease phenotypes from free clinical text. Moreover, researchers have recently used GPT (Generative Pre-trained Transformer) [22], a transformer-based model specifically crafted for language generation, with the more powerful releases being proprietary to OpenAI. In our previous work [27], we evaluated ChatGPT off-the-shelf model for phenotypes extraction, which happen to underperform compared to PhenoBERT. This confirmed that LLM need to be customized (through fine-tuning or prompting) or combine with other technics to reach better performance. In this spirit, PhenoBCBERT [6] and PhenoGPT [6] models leverage large language models to automate the detection of phenotype terms, including those not in the current HPO. These models recently improved the state-of-the-art results for phenotypes extraction. However, the exact contribution of the LLM steps as well as their performance with implicit phenotypes have not been investigated so far.

C. Evaluation data and metrics

Assessing a model's performances for BEL requires manually-annotated data by experts, and checking how the model's predictions match the labels. By comparing the model's predictions against the ground truth annotations, metrics like Precision, Recall, and F1 scores respectively quantify how accurate the model is, how well it finds the right spans, and how it balances precision and recall.

Evaluation is a key step as it determines the model performances on real-world data, and allows a fair comparison between systems. Qualitative analysis of the evaluation results can also gives researchers clues to improve their solutions. For phenotypes extraction, the academic community used mainly the GSC+ dataset [7], comprising the abstracts of 228 disease research articles, and the ID-68 dataset [8] featuring 68 authentic clinical notes from about families with intellectual disabilities. In both datasets, the phenotypic descriptions in each patient family's clinical notes were manually annotated with HPO terms, to assess and comprehend a model's capability in accurately classifying, capturing context, and maintaining a balance between false positives and false negatives. However, these two dataset do not necessarily cover all the real world clinical reports types we may encounter. In particular, we found many situations where phenotypes are more implicitly referenced. In order to assess in-depth performances of the solutions, we generated an internal corpus called *CHU-50*. It comprises 50 medical reports, each written in free text and in French by several doctors specialized in clinical genetics (from the Clinical Genetics Department of the University Hospital of Rennes) and corresponding to reports that may be produced after an initial clinical genetics consultation for evaluation of potential intellectual disability and/or poly-malformative syndrome. These reports have been translated into English to fairly evaluate solutions that do not support French. We plan to release this dataset in future publications.

As we are dealing with a set of reports within a corpus, the previous metrics can be computed at report-level or at corpus level, which has been defined as macro and micro-average approaches:

- Macro-average Precision (Per Report) is calculated by considering the precision of each individual report (document) independently then averaged across all reports. This provides an average precision score that treats each report equally, regardless of class distribution.
- Micro-average Precision (Across All Reports) is calculated by aggregating the true positives, false positives, and false negatives across all reports and then computing the overall precision. This metric treats each prediction (or sample) equally across all reports, providing an overall precision score for the entire dataset.

And conversely for Recall and F1 scores.

III. Analysis

We will analyse and build upon the recent breakthrough in biomedical text processing, particularly focusing on the integration of deep learning and transformer-based models and large language models. Given the efficiency demonstrated by PhenoBCBERT, PhenoGPT, and the hybrid approach of PhenoBERT, our research focuses on these solutions. By doing so, we aim to contribute to the ongoing progress in biomedical informatics, with a specific emphasis on enhancing the recall of phenotype term detection in clinical settings. In fact, as we plan to develop a tool to help clinicians to annotate reports without missing any relevant phenotype that could improve diagnosis, this metric is preponderant.

A. PhenoBERT Workflow

As PhenoBERT stands out as a premier open-source solution, our analysis centers around it, aiming to illuminate both its strengths and weaknesses

In the PhenoBERT study, researchers aimed to develop an effective hybrid method for identifying HPO terms in clinically relevant text segments (CTSs). The resulting model was compared with various methods, including dictionary-based and other deep learning-based approaches. PhenoBERT consistently outperformed all competing methods, with a more noticeable advantage in challenging phenotype extraction tasks. Furthermore, PhenoBERT exhibited a fourfold increase in speed compared to PhenoTagger due to the use of two layers of CNNs for pre-selection.

To identify a broad range of phenotypes within the unstructured clinical text, PhenoBERT authors devised a system incorporating multiple layers of integrated algorithms, such as Bi-LSTM, string-matching, CNNs, and BERT. The outlined workflow in the original paper comprises two primary processes as shown in Figure 2: text segmentation and concept classification.

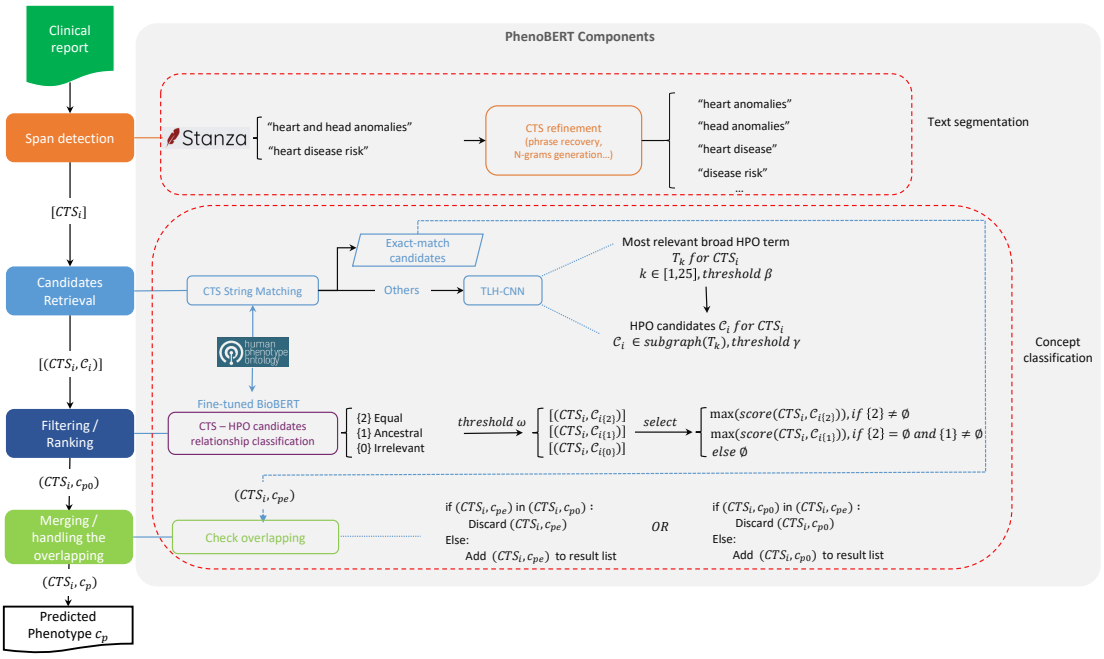

Fig. 2. Our analysis of the PhenoBERT workflow

Text segmentation: This process primarily utilizes a deep learning technique (Bi-LSTM) known as Stanza [15]. Stanza is applied on the clinical free text to segment it into sentences and then identify CTSs within each sentence using its "ner-i2b2" processor and the "mimic" package. To capture additional segments possibly missed by Stanza, each sentence undergoes further segmentation using conjunctions and punctuation defined in the NLTK library. Segments not overlapping with those identified by Stanza are considered additional CTSs. The refinement procedure for all identified CTSs involves three key steps: phrase recovering, stop words filtering, and n-grams extraction, each contributing to a more nuanced understanding of the input text. Overall, this segmentation process is designed to meticulously improve the identification and extraction of clinically relevant information from the input text, ensuring a thorough analysis of the content.

Concept classification: This process involves several steps to achieve the final result, including string matching, deep learning methods, and post-processing to handle overlapping CTSs.

Dictionary-based Matching: The system uses dictionary-based string matching to check if a CTS matches any terms or synonyms in the HPO dictionary. This Matched (c_{pe}) will be added to a temporary result list with a score of 1; non-matched (CTS_i) will proceed to the next step.

Hierarchical CNNs and BERT: This step combines Two-Level Hierarchical CNNs (TLH-CNNs) and BERT to determine the most relevant HPO terms. The first-level CNN classifies CTS_i into 25 broad HPO subgroups based on a threshold β (0.6), then the second-level CNN refines the classification to produce a list of candidate HPO terms based on a threshold γ (0.8). If no candidate HPO terms exceed the respective thresholds β or γ, the CTS is discarded. Then BERT will evaluate each pair formed by the CTS and a candidate HPO term $[(CTS_i, C_i)]$ to determine the most relevant one with prediction scores exceeding a threshold ω (0,9) and consider it as a final matched HPO term c_{p0} . The selected pairs are divided into two groups according to their labels: those with label 2 (equal relationship) $[(CTS_i, C_{i\{2\}})]$ and those with label 1 (ancestral relationship) $[(CTS_i, C_{i\{1\}})]$. If the label 2 group exists, then the HPO term $C_{i\{2\}}$corresponding to the pair with the greatest score is assigned to the CTS_i. Otherwise, the HPO term $C_{i\{1\}}$ corresponding to the top-ranked pair in the label 1 group is assigned to the CTS_i. If neither group exists, the CTS_i is discarded.

Handling Overlapping: This step handles overlapping CTSs, acknowledges the possibility of overlapping CTSs that are assigned to the same HPO term due to n-gram extraction in text segmentation, and resolves conflicts by favoring the longer CTS since it typically represents a more specific phenotypic description.

For instance, in cases where "peripheral neuropathy" and "neuropathy" are overlapping CTSs, the longer one ("peripheral neuropathy") would be favored as it likely pertains to a more specific HPO term.

The analysis of the PhenoBERT results showed that while PhenoBERT is highly effective for mining clinical text data efficiently, it struggles to account for the context associated with specific phenotypes. For example, a trait might be mentioned as absent, such as 'no autism,' or it might refer to a patient's family member rather than the patient. In these cases, PhenoBERT cannot include this contextual information

in its output, and a deeper analysis of the results revealed that 60% of the outcomes on GSC+ and 71% on ID-68 were achieved through the initial dictionary-based string matching step. The number of candidates generated by Stanza significantly contributed to these successful results. This finding led us to experiment with replacing Stanza with other available technical alternatives to further improve performance.

B. PhenoBCBERT and PhenoGPT Workflows

In the study [6], researchers utilized large language models to improve phenotype recognition by proposing two transformer-based models: PhenoBCBERT (based on BERT) and PhenoGPT (based on GPT). Both models require accurately labeled data containing phenotypic information. However, since GPT models are pre-trained on much larger datasets compared to BERT, they need a relatively smaller fine-tuning dataset to achieve similar results. Due to the fundamental differences in their structures, different labeling strategies were used to train the PhenoBCBERT and PhenoGPT models for phenotype entity recognition.

For PhenoBCBERT, they initially trained a phenotype recognition model on top of Bio+ClinicalBERT for rare disease-specific NLP text mining tasks. PhenoBCBERT was developed by initializing from the Bio+Clinical BERT model, which was then fine-tuned on a mixed-supervised dataset consisting of 3,400 automatically labeled clinical notes (using PhenoTagger) and 460 hand-labeled clinical notes from their in-house dataset, allowing the recognition of terms outside the standard HPO vocabulary. For PhenoGPT, they used a range of GPT models, including GPT-J-6B, Falcon-7B, and LLaMA-7B for open-source versions, and GPT-3 for the closed-source version. Both versions were fine-tuned using the public BiolarkGSC+ dataset. Instead of training a GPT model for named entity recognition (NER), they labeled the training data to match the model's nature as a generative decoder. For a given clinical abstract, they generated text by appending phenotype entities with their associated HPO IDs to the abstract for either prompt-based learning or fine-tuning. PhenoBCBERT and PhenoGPT models could identify a wider range of phenotype concepts, also show strong performance in case studies on biomedical literature.

Inspired by these findings, we built experiments leveraging the Large Language Models (LLMs) paradigm to enhance phenotype recognition.

C. LLM Alternative

In order to go one step further in phenotype recognition field, we propose to use LLM as an alternative for Span Detection and optionally for Entity Linking to evaluate if an improvement can be reached. As PhenoBERT is the only state-of-the-art solution released in open source, we focus our LLM alternative comparison towards this solution. For transparency and reproducibility purposes, the optimal prompt used in our experiment is given in the supplementary material section.

The main experiment (PhenoBERT$_{LLMspan}$) consists in replacing the Stanza module by a LLM (ChatGPT-3.5). We first tokenized the reports into sentences, and query the LLM for each one as previous work showed that sentence-level strategy gives better results than report-level one [27].

We designed a span detection prompt leveraging the In-Context Learning (ICL) ability to guide the generation with few examples (not included in the evaluation set). The goal of this experiment is to assess the added value of a LLM for this first critical step. As our previous analysis highlighted the fact that a majority of phenotypes from ID-68 and GSC+ were detected with string matching, we also run an evaluation based on an internal dataset (CHU-50) with a higher ratio of implicit phenotypes. To that end, in addition to manual anonymization and random sentence-by-sentence query strategy, we benefited from access to an Europe-located private OpenAI server to avoid data leakage.

In particular, the experiments show interesting insights regarding the public dataset Ground Truth, as presented in the following section.

IV. Results

A. Quantitative Results

When LLM is used as a span detection module within PhenoBERT workflow, we notice a slight improvement on the ID-68 dataset (Table I), more noticeable for the macro-average scores. This marginal improvement is expected knowing the nature of the dataset, where phenotypes are explicitly formulated.

System	Micro-Average			Macro-Average		
	P	R	F1	P	R	F1
PhenoBERT	93.98	78.12	85.32	94.47	77.56	85.18
PhenoBERT$_{LLMspan}$	93.85	**78.25**	**85.34**	**94.77**	**78.86**	**86.09**

P=Precision, R=Recall, F1=F1-score

TABLE I
ID-68 Results[2]

On the GSC+ dataset, the precision remains similar to PhenoBERT but the recall drops slightly. Our hypothesis to explain this lower recall is that we designed our ICL prompt with examples similar to the ID-68 sentences as we first focused on this dataset. We plan to run a new experiment with ICL examples fitting the GSC+ sentences to see if it affects the results.

System	Micro-Average			Macro-Average		
	P	R	F1	P	R	F1
PhenoBERT	79.96	66.98	72.90	78.98	70.83	74.68
PhenoBERT$_{LLMspan}$	79.73	61.65	69.54	**79.19**	66.69	72.41

P=Precision, R=Recall, F1=F1-score

TABLE II
GSC+ Results[2]

[2]PhenoBERT scores obtained by reproducing the experiment.

Table III shows that the overall performance of PhenoBERT is much lower on the more challenging CHU-50 dataset compared to ID-68 and GSC+. Both precision and recall are dropping, the latter showing an even more significant drop. This is expected as we are now dealing with more implicit phenotypes. When using LLM span detection module, recall increases slightly, but at the cost of a lower precision.

System	Micro-Average			Macro-Average		
	P	R	F1	P	R	F1
PhenoBERT	**62.74**	39.34	48.35	**63.28**	38.66	48.00
PhenoBERT$_{LLMspan}$	58.36	**42.66**	**49.29**	58.09	**41.78**	**48.60**

P=Precision, R=Recall, F1=F1-score

TABLE III
PHENOBERT PERFORMANCE ON CHU-50

The added value of LLM for span detection is not obvious as it seems to bring marginal improvement on recall. However, as the scores are computed at the end of the PhenoBERT pipeline, the real added value of the LLM is not clear as relevant spans may have been filtered by the internal classification processing. Hence, in order to get more clues on its relative performance, we performed qualitative analysis on different reports, focusing on the output of the span detection step.

B. Qualitative Results

To get more insights, we compared the output of raw text segmentation and CTS refinement of the two pipelines. We noticed that Stanza often misses relevant semantic information (e.g. adjective, negation) as well as medical tokens, while the LLM identifies much more elements in the sentence. As shown in Table IV, *barely can feel* and *no* are important information to associate a negation to the output phenotype. The acronym *TSH* is also essential in the second sentence, but missed by Stanza.

Figure 3 shows an example of the extracted phenotypes using the original PhenoBERT (with Stanza) compared to the ones output by our proposed method on a sentence from the ID-68 dataset. As we can see, the better LLM span detection results in two additional correct phenotypes prediction compared to the original PhenoBERT. It worth noting that the LLM span detection module was actually able to detect *delay in fine motor and language domains* as well as *immature finger grasp*, but no related phenotypes were found by the PhenoBERT classification stage.

In fact, as the LLM is used to replace Stanza only, its output goes into the same refinement process resulting in generating different n-grams from individual tokens. Doing so, the additional information brought by the LLM is often discarded, and the most explicit n-grams have statistically a higher probability to be kept in the classification filtering step. This may explain the marginal added value of introducing LLM for span detection especially with explicit dataset like ID-68.

In addition, analysis on our internal dataset revealed that the original PhenoBERT missed a lot of annotated spans, whereas our LLM span detector module detected more spans than the Ground Truth. This result explains the low recall from PhenoBERT on the CHU-50 dataset, and may also suggest that the LLM identifies many spans (without sharp discrimination), hence improves recall at the cost of precision.

However, when analyzing the detected spans related to out-of-GT predicted phenotypes, our clinicians noticed that the majority of them were actually relevant. In other words, the LLM was able to detect spans missed by the human annotators. The same observations apply to the ID-68 dataset. Table V presents some out-of-GT spans detected by the LLM that have been considered as relevant by clinicians, which turned out to be the case for 79% of this type of spans for the ID-68 dataset. This is an important observation, as it means that even if the existing dataset are essential pillars for systems evaluation, we can not rely blindly on the annotations as they are not exhaustive and may result in performance biases.

As shown in Table V, the LLM may reformulate the spans from time to time when the detected information is not contiguous (e.g.: *"IQ 20-30"* span is related to *"IQ was estimated to be 20-30"*). This can be challenging for evaluation as we loose the span indices, and some update need to be done to take into account this new approach.

V. CONCLUSION

Important improvements have been achieved in the last few years regarding phenotypes extraction from text, especially with the use of LMs combined with dedicated techniques to cover the different steps involved: text segmentation to identify relevant spans, and candidates retrieval and classification to select the most probable phenotypes related to these spans. However, our analysis of PhenoBERT, one of the best existing solution, revealed some limitations: relevant spans are not always detected, and the inner processing favors explicit matching, leading to significant performance drop when dealing with more realistic datasets having implicit references to phenotypes. Integrating a new span detector component based on LLM improves the results, more obviously observed for the recall, but the improvement remains marginal for the existing dataset with explicit annotations. Qualitative analysis reveals that our LLM span detector is much better than the PhenoBERT one, but this added value is largely erased during the following filtering stages. In addition, using LLM for span detection led to an insightful discovery: a majority of out-of-GT spans output by the LLM were actually relevant. On the positive side, this opens an interesting opportunity to use LLM to identify more spans than human annotators, which can help reduce such work time, while improving deep phenotyping. On the downside, it shows that current dataset annotations happen to be non exhaustive, leading to a potential evaluation bias. In order to improve the evaluation process for future work, we believe augmented versions of current dataset such as ID-68 and GSC+ could be generated leveraging LLM for span detection. More implicit dataset could also be a great asset to help the community in finding better solutions. We

Sentence	Stanza span output	LLM span output	Target phenotype
Her parents report that she barely can feel pain, and she has no tears when she cries even though she has normal sweating.	pain tears	barely can feel pain no tears	Pain HP:0012531 (Neg) Epiphora HP:0009926 (Neg)
His TSH has been persistently mildly elevated.	persistently mildly elevated	TSH has been persistently mildly elevated	Elevated circulating thyroid-stimulating hormone concentration HP:0002925

TABLE IV
SPAN OUTPUT COMPARISON (ID-68 SENTENCES)

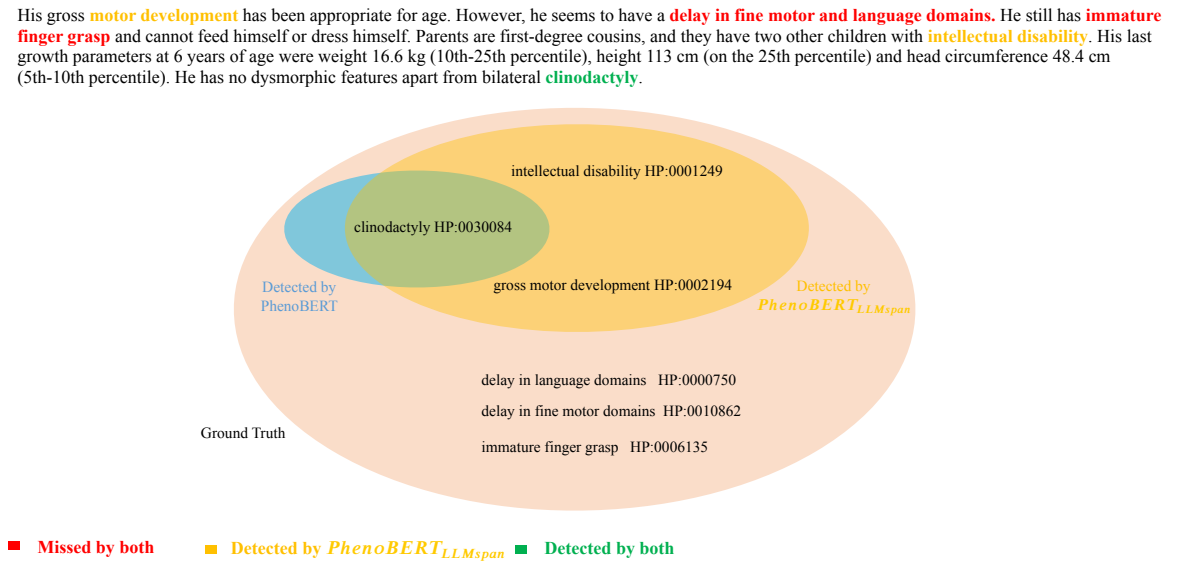

Fig. 3. Phenotypes Extraction in PhenoBERT and PhenoBERT$_{LLMspan}$ (ID-68 sample)

Sentence	Out-of-GT span	Is span relevant?†
After birth, she was noted to be floppy with poor sucking.	floppy	yes
Her seizures semiology appears to be as opisthotonic posturing.	opisthotonic posturing	yes
Currently, she has a developmental delay affecting all domains, and her IQ was estimated to be 20-30.	IQ 20-30	yes
She can vocalize, but cannot talk.	cannot talk	yes
Her growth parameters at 7 years of age were weight 12.7 kg (<3rd percentile)	3rd percentile weight	yes
Developmentally, he started to say dada and mama at 4 years of age.	developmental delay	yes
She is still unable to sit	unable to sit	yes
She was noted to have aggressive behavior that led to drug treatment.	drug treatment	no
A 3 years old boy who was born at 34 weeks gestation via an emergency C-Section due to fetal distress.	34 weeks gestation	no

†Manually curated by clinicians

TABLE V
EXAMPLES OF OUT-OF-GT DETECTED SPANS ON ID-68

hope to be able to provide such kind of dataset in the coming months.

Overall, this analysis suggests to find new approaches for candidates selection from high-quality spans, based on the LLM paradigm. In fact, as our study was limited to the very first step of phenotypes extraction, more work has to be done on the candidates selection step in order to identify bottlenecks and underlying challenges. In particular, we currently evaluate a full LLM pipeline, and plan to identify the relative limitations of new approches such as Retrieval Augmented Generation (RAG) [28]. It worth mentioning that some new evaluation protocol has to be implemented to take into account the fact that output spans can be reformulated, leading to indexes discrepancies. We are convinced that current state-of-the-art can be greatly improved in order to further reduce the diagnostic wandering.

REFERENCES

[1] French, E., & McInnes, B. T. (2023). An overview of biomedical entity linking throughout the years. Journal of Biomedical Informatics, 137, 104252–104252. https://doi.org/10.1016/j.jbi.2022.104252

[2] Köhler, S., et al. (2017). "The Human Phenotype Ontology in 2017." Nucleic Acids Research, 45(D1), D865-D876. DOI: 10.1093/nar/gkw1039.

[3] Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70. IEEE.

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[5] Feng Y, Qi L, Tian W. PhenoBERT: A Combined Deep Learning Method for Automated Recognition of Human Phenotype Ontology. IEEE/ACM Trans Comput Biol Bioinform. 2023 Mar-Apr;20(2):1269-1277. doi: 10.1109/TCBB.2022.3170301. Epub 2023 Apr 3. PMID: 35471885.

[6] Yang J, Liu C, Deng W, Wu D, Weng C, Zhou Y, Wang K. Enhancing Phenotype Recognition in Clinical Notes Using Large Language Models: PhenoBCBERT and PhenoGPT. ArXiv [Preprint]. 2023 Nov 9:arXiv:2308.06294v2. Update in: Patterns (N Y). 2023 Dec 05;5(1):100887. PMID: 37986722; PMCID: PMC10659449.

[7] M. Lobo, A. Lamurias, and F. M. Couto, "Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules," Biomed Research International, vol. 2017, 2017

[8] Anazi S, Maddirevula S, Salpietro V, Asi YT, Alsahli S, Alhashem A, Shamseldin HE, AlZahrani F, Patel N, Ibrahim N, Abdulwahab FM, Hashem M, Alhashmi N, Al Murshedi F, Al Kindy A, Alshaer A, Rumayyan A, Al Tala S, Kurdi W, Alsaman A, Alasmari A, Banu S, Sultan T, Saleh MM, Alkuraya H, Salih MA, Aldhalaan H, Ben-Omran T, Al Musafri F, Ali R, Suleiman J, Tabarki B, El-Hattab AW, Bupp C, Alfadhel M, Al Tassan N, Monies D, Arold ST, Abouelhoda M, Lashley T, Houlden H, Faqeih E, Alkuraya FS. Expanding the genetic heterogeneity of intellectual disability. Hum Genet. 2017 Nov;136(11-12):1419-1429. doi: 10.1007/s00439-017-1843-2. Epub 2017 Sep 22. Erratum in: Hum Genet. 2017 Dec 29;: PMID: 28940097.

[9] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36. doi: 10.1136/jamia.2009.002733. PMID: 20442139; PMCID: PMC2995713.

[10] Clement Jonquet, Nigam Haresh Shah, Cherie Youn, Mark A. Musen, Chris Callendar, Margaret-Anne D. Storey. "NCBO Annotator: Semantic Annotation of Biomedical Data" (2009). Available at: https://api.semanticscholar.org/CorpusID:6789996

[11] Smith,B. Ashburner,M., Rosse,C. et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol., 25, 1251–1255.

[12] Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, Genetti CA, Brownstein CA, Schmitz-Abe K, Schoch K, Cope H, Signer R; Undiagnosed Diseases Network, Martinez-Agosto JA, Shashi V, Beggs AH, Wheeler MT, Bernstein JA, and Bejerano G (2018). ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genetics in Medicine, 2018.

[13] Clement Jonquet, Nigam Haresh Shah, Cherie Youn, Mark A. Musen, Chris Callendar, Margaret-Anne D. Storey. 'NCBO Annotator: Semantic Annotation of Biomedical Data' (2009). Available at: https://api.semanticscholar.org/CorpusID:6789996

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019, arXiv:1810.04805 [cs.CL].

[15] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," 2020, arXiv:2003.07082 [cs.CL].

[16] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans (Eds.), Doha, Qatar, Oct. 2014, Association for Computational Linguistics, pp. 1532–1543. https://aclanthology.org/D14-1162, doi: 10.3115/v1/D14-1162.

[17] Kaj Bostrom and Greg Durrett. "Byte Pair Encoding is Suboptimal for Language Model Pretraining," arXiv preprint arXiv:2004.03720 (2020), cs.CL.

[18] Keiron O'Shea and Ryan Nash. (2015). An Introduction to Convolutional Neural Networks. arXiv preprint arXiv:1511.08458.

[19] Robin M. Schmidt. (2019). Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. arXiv preprint arXiv:1912.05911.

[20] Arbabi A, Adams DR, Fidler S, Brudno M. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. JMIR Med Inform. 2019 May 10;7(2):e12596. doi: 10.2196/12596. PMID: 31094361; PMCID: PMC6533869.

[21] E. G. Piotr Bojanowski, Armand Joulin, Tomas Mikolov, "Enriching Word Vectors with Subword Information," arXiv preprint arXiv:1607.04606, 2016

[22] Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. (2023). Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. arXiv preprint arXiv:2305.10435.

[23] Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, Zhiyong Lu, PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology, Bioinformatics, Volume 37, Issue 13, July 2021, Pages 1884–1890, https://doi.org/10.1093/bioinformatics/btab019

[24] Hochreiter, Sepp , Schmidhuber, JÃ¼rgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[25] Steven Bird, Ewan Klein, and Edward Loper. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

[26] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini and Hervé Jégou. (2024). The Faiss library.

[27] Thomas Labbé, Pierre Castel, Jean-Michel Sanner and Majd Saleh. (2023). ChatGPT for phenotypes extraction: one model to rule them all? 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).

[28] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., & Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

VI. Supplementary material

We tested several prompts to query the LLM to detect spans related to phenotypes. The following prompt corresponds to the optimized one used in our experiment. We provide the ICL examles template without the real examples as some of them are inspired from our internal dataset. Eight ICL examples were provided covering several complexity levels (adding more examples marginally improved the results, while being more expensive).

> You are an experimented clinician with an exhaustive knowledge of human phenotypes ontology. Given a sentence, you must identify the spans related to possible phenotypes, either explicitly or implicitly. You should keep in the span all words related to the phenotype that should be informative (such as negation or adjective). You may reformulate the span if needed. If you don't detect any span or if you don't know, don't try to make up an answer, just write 'None'.
>
> SENTENCE: [example sentence].
> ==========
> Span: [list of spans related to phenotypes in the given sentence]

We plan to provide the full prompt with synthetic examples in our next publication.