ChatGPT for phenotypes extraction: one model to rule them all?

Thomas Labbé, Pierre Castel, Jean-Michel Sanner, Majd Saleh

ChatGPT for phenotypes extraction: one model to rule them all?

Thomas Labbé[1,2] Pierre Castel[2] Jean-Michel Sanner[1,2] Majd Saleh[2]

***Abstract*— Information Extraction (IE) is a core task in Natural Language Processing (NLP) where the objective is to identify factual knowledge in textual documents (often unstructured), and feed downstream use cases with the resulting output. In genomic medicine for instance, being able to extract the most precise list of phenotypes associated to a patient allows to improve genetic disease diagnostic, which represents a vital step in the modern deep phenotyping approach. As most of the phenotypic information lies in clinical reports, the challenge is to build an IE pipeline to automatically recognize phenotype concepts from free-text notes. A new machine learning paradigm around large language models (LLM) has given rise of an increasing number of academic works on this topic lately, where sophisticated combinations of different technics have been employed to improve the phenotypes extraction accuracy. Even more recently released, the ChatGPT[1] application nevertheless raises the question of the relevance of these approches compared to this new generic one based on an instruction-oriented LLM. In this paper, we propose a rigorous evaluation of ChatGPT and the current state-of-the-art solutions on this specific task, and discuss the possible impacts and the technical evolutions to consider in the medical domain.**

***Clinical relevance*— Deep phenotyping on electronic health records has proven its ability to improve genetic diagnosis by clinical exomes [10]. Thus, comparing state-of-the-art solutions in order to derive insights and improving research paths is essential.**

I. INTRODUCTION

Data powerness, boosted during the last decade thanks to machine learning capabilities, is based on the fact that data brings value to those who know how to make it speak. If correlation or pattern recognition were the first obvious applications based on structured or semi-structured data, deriving insights from unstructured ones such as human-generated text became rapidly a new challenge, tackled in the Natural Language Processing (NLP) field. Text processing can be roughly summarized in 2 steps: first, compute a numerical representation of the text (statistical or probabilistic); second, use this representation to process downstream tasks (classification, translation, generation). One of the most interesting tasks for many domains is to extract factual knowledge from unstructured text, also known as Information Extraction (IE). In medicine, the target knowledge can be symptoms, drugs, phenotypes or any others characteristics associated to a patient that can ease doctors in making decisions. Hence, finding solutions to extract such knowledge is vital to improve patients care. Nonetheless, this represents a challenging task given the wide variety of language forms and domain-specific terms as well as implicit meanings which can hardly be understood by non-experts. That is why the quality of the text representation (step 1) is crucial to have good results in the extraction task (step 2), and this is where recent advances in language modeling have become valuable assets. Indeed, if the first works in this area were based on rule-based parsers, the most recent ones significantly improved the results leveraging modern pretrained Transformer models. Nonetheless, relying solely on such generic representation models has not proven to be sufficient, and instead several mechanisms were combined together to adapt it (e.g. domain adaptation and fine-tuning) and bypass its weaknesses (e.g. span detection, dictionaries). Then came ChatGPT, an application based on a large language model (LLM) able to answer many instructions without being specifically trained on domain-oriented tasks, and new research questions arose:

1) What is the performance of ChatGPT on IE, and especially on phenotypes recognition?
2) How does it compare to the current state-of-the-art solutions?
3) To what extent very large language models may become a key component for computational medicine?

In this paper, we will focus on phenotypes extraction from text in the deep phenotyping perspective, where the objective is to associate a list of phenotypes to a given clinical report, using Human Phenotype Ontology (HPO)[11] as a target reference.

II. RELATED WORK

A. Information Extraction

First approaches for recognition of HPO terms in medical reports were mainly based on dictionary string matching parsers. For instance, Open Biomedical Annotator[5] represents a string matching-based tool for extracting HPO terms. The list of direct HPO concept provided can then be expanded using a semantic annotation component. MetaMapLite [18] and Clinphen[17] represent other examples of rule-based methods. They all used a dictionary built from the HPO database to perform a rule-based process analysis.

Machine learning approaches were also tested for biomedical IE, such as a combined LSTM - CRF model trained to recognize five entities belonging to a biomedical domain [7]. A novel RNN architecture trained with a corpus of weakly labelled data to predict a larger number of concepts from the international classification of diseases (ICD) in medical notes [15] is an other example.

However, annotation of a medical corpus with HPO terms is more difficult than classical IE classification tasks due to the number of HPO classes to be learned (more than 14 000 HPO terms), and also due to the scarcity of public data available. This is a challenging aspect of this particular task for

[1]Orange Labs, Rennes, France thomas.labbe@orange.com
[2]b<>com Research Institute, France majd.saleh@bcom.com
[1]https://chat.openai.com/

machine learning classification, and alternative approaches have been proposed.

Neural Concept Recogniser [2], uses deep learning blocks to find HPO concepts in medical reports. This approach is based on similarity measurement between a vector representation of the HPO names and text tokens vectors representation based on fast-text [3]. Sequence tokens vectors representation are learned with a simple convolutional encoder. Neural Concept Recogniser did not use any rule-based function and is only based on semantic similarity. When released, Neural Concept Recogniser outperformed the state-of-the-art of information extraction of HPO terms from medical texts. The growing success of these deep learning approaches and the wide dissemination of Transformer-based models from 2019 unsurprisingly paved the way for the use of language models for this extraction task.

B. Language Models Era

Performances in many NLP tasks and among them IE has then been strongly enhanced by the development of learned language models.

Phenotagger [12], published in 2021, uses a pretrained BioBERT [9] language model fine-tuned on an augmented data set built with texts attached to HPO terms. A classification process is then used on a sliding window composed of two to ten tokens. The selection of the HPO terms combined a dictionary matching method on HPO id with the classification results to select the most relevant phenotypes.

BERT [8] language model was also leveraged by PhenoBert [16], a solution released in 2022 and based on a sophisticated pipeline composed of several blocks to identify automatically some of the 14000 HPO terms. The first step consisted in applying a deep learning based method [13] to select clinically relevant text segments. Then a dictionary rule-based method is used trying to extract explicit HPO terms. Next, for text spans where no HPO terms were found, a two levels of 26 Convolutional Neural Network (CNN) classification process was applied. The first CNN level tried to find the most relevant of the 25 children concepts at the root of the HPO tree which could relate to the text span. The second CNNs classification level tried to find a list of children candidates for each text span. The best terms were then selected using a BERT based similarity measurement on embedded sentences pairs composed of the text span on the one hand, and the HPO term itself on the other hand. PhenoBERT outperformed the state of the art represented by PhenoTagger.

More recently, the LLM field was shaken up with the release of ChatGPT, a language model optimized for interactive dialogue which provides impressively realistic answers to human questions, in a wide variety of domains. ChatGPT is fine-tuned from GPT-3.5 pretrained model (which can be considered as an improved version of GPT-3 [4]), using supervised reinforcement learning from human feedback (RLHF). This last approach is the basis of InstructGPT [14], a previously released prompt-oriented model and which can be considered to be a ChatGPT sibling. Without being specifically trained on medical domain and tasks, it is able to extract phenotypes concepts with corresponding HPO ID and label from any text.

Considering its performance in a wide spectrum of tasks, and given first conclusive experiments on a few clinical notes samples, it was legitimate to question its relevance by evaluating it on reference datasets.

III. METHODOLOGY

In order to fairly compare ChatGPT with state-of-the-art approaches, we implemented a methodology taking into account the prompt-oriented nature of this model.

The OpenAI API[2] was used to access the model. As a generative model, ChatGPT needs a first prompt context which will guide it to build a relevant answer. In that sense, it is particularly sensitive to the formulation of the prompt. In our case, the prompt is composed of an intent following by the text to extract the phenotypes from. Regardless of the performance, we first tested several intent prompts in order to find ones that fit the purpose. The mandatory requirement was to get the exact HPO labels and the corresponding IDs, the latter being the unambiguous keys to make the evaluation: a candidate will be considered incorrect as soon as its ID does not match the target one, regardless of the distance of candidate and target in the ontology tree. Moreover, we wanted to assess the capability of ChatGPT to provide a confidence score associated to the extracted phenotypes, so it can be used for further threshold or top-N filtering strategies. In order to evaluate the influence of the prompt, two slightly different intent formulations were tested (Fig. 1).

Prompt#1	"Are there HPO phenotypes associated with this text (if so write in the form [HPO labels | HPO IDs | confidence score between 0 and 1]): "
Prompt#2	"Extract HPO phenotypes from this text (parsed as a list in brackets [HPO name | HPO ID]): "
Target text	"*A decrease in fetal movements was noted during pregnancy, and her birth weight was 2.8 kg*."
Answer#1	[Decreased fetal movement | HP:0001558 | 0.9], [Low birth weight | HP:0001518 | 0.8]
Answer#2	[Decreased fetal movement | HP:0001558], [Low birth weight | HP:0001518]

Fig. 1: API Intent Prompts

In this example, ChatGPT seems to have a good knowledge of the HPO ontology. Nonetheless, we noticed that the label-ID association was often spurious. This was confirmed by some in-depth tests such as the one presented in Fig. 2.

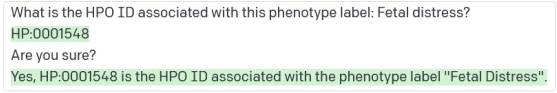

Fig. 2: Spurious HPO Label - ID Association

We shall specify that ChatGPT integrates an active learning mechanism that enables the model to use human feedback and correct inaccurate answers accordingly[3]. This means the previous example might become correct if someone points out this wrong answer. However, this underlines the fact that it is not possible to rely on these associations without reservation. As the scores are computed through IDs

[2]https://openai.com/api/

[3]https://help.openai.com/en/articles/5722486

comparison (predicted vs Ground Truth), their relevance is crucial. Hence, we decided to evaluate 3 label-ID association methods:

- Raw: the Label-ID pairs returned by ChatGPT;
- PhenoB: from label returned by ChatGPT, use the `get_most_related_HPO_term()` function from PhenoBERT[4] which is a semantic search method based on FastText model to get the most relevant ID from a string;
- HPOapi: from a label returned by ChatGPT, use the HPO API[5] to get the corresponding ID.

The objective of phenotypes extraction is to associate a list of phenotypes to a given clinical report. However, two text-levels could be considered to process the extraction:

- Sentence-level: the text is first tokenized into sentences, then each sentences is sent to ChatGPT, and the generated lists of phenotypes are concatenated into a single list (with deduplication) which is used for evaluation;
- Report-level: the whole report text is sent to the ChatGPT API, which generates a single list of phenotypes.

The choice of the best approach is not trivial, as the text length is known to have high influence upon LLM results. To further evaluate the capabilities of ChatGPT, we choose to run both approaches with different parameters.
In order to fairly compare ChatGPT results with others state-of-the-art solutions, we used the evaluator developed by the authors of PhenoBERT [16]. The evaluator computes the following metrics: precision, recall and F1-score, averaged at the document level (micro), and at the entire corpus level (macro).

IV. EXPERIMENTS

A. Dataset and Experimental Settings

Two publically available labeled datasets have been used: GSC+ [13] and ID-68 [1]. The Gold Standardized Corpora (GSC+) dataset consists of 228 diseases research articles abstracts annotated with HPO terms by experts. The annotation includes the HPO term, the position of the annotation in the text segment, and the considered text segment itself. The ID-68 dataset consists of 68 clinical notes from families with intellectual disabilities annotated by experts in the same way as in GSC+.

Several parameters can be set in the request to the OpenAI API. Apart from the model itself which is mandatory, the temperature parameter is important to consider, as it represents the degree of randomness for the text generation: higher temperature encourages the model to be more creative, which means answers to the very same prompt can change over time. In order to make the experiment reproducible, we first set a temperature to 0 for the two target datasets. Then, we ran other experiments on a single dataset with a temperature set to 0.7 (default value of ChatGPT) to analyze the influence of this parameter, knowing that in our case (and more broadly in medical domain), the ideal temperature should be 0 to maximize exactness. Finally, we set the maximum tokens parameter to 80 for the sentence-level approach (allowing the model to generate up to 6 phenotypes by sentence), and to 256 for the report-level one.

The experiments and the associated parameters are summarized in Table. I.

Expe	Description	Target Dataset	Parameters
#1	Baseline	ID-68 and GSC+	model: text-davinci-003, prompt: Prompt#1, Temperature: 0, Max_tokens: 80 for sentence-level / 256 for report-level, Label-ID association: {Raw, PhenoB, HPOapi}
#2	Prompt influence	ID-68	model: text-davinci-003, prompt: {Prompt#1, Prompt#2}, Temperature: 0, Max_tokens: 80 for sentence-level / 256 for report-level, Label-ID association: HPOapi
#3	Temperature influence	ID-68	model: text-davinci-003, prompt: Prompt#2, Temperature: {0, 0.7}, Max_tokens: 80 for sentence-level / 256 for report-level, Label-ID association: HPOapi

TABLE I: Experimental Settings

B. Results and Discussion

Results of the first experiment are shown in Tables II and III. Using the ChatGPT response with the raw Label-ID association leads to very low scores. The spurious IDs obviously cause performance degradation as the results are significantly better when IDs are retrieved with alternative offline methods. The PhenoBERT semantic search method gives the best results, highlighting the fact that such ranking mechanism add value to any candidates generation approach. Interestingly, the sentence level method outperforms compared to report level except for precision metric. An hypothesis to explain this point is that in sentence level strategy, ChatGPT returns much more candidates (concatenation of a maximum of 80 tokens per sentence vs 256 for a whole report), which logically reduces precision.

	System	Micro-Average			Macro-Average		
		P	R	F1	P	R	F1
	Clinphen	0.74	0.61	0.67	0.74	0.61	0.67
	MetaMap Lite	0.80	0.59	0.68	0.81	0.59	0.68
	PhenoTagger	0.89	0.75	0.82	0.89	0.76	0.82
	PhenoBERT	**0.94**	**0.78**	**0.85**	**0.94**	**0.77**	**0.85**
Sentence	ChatGPT_raw	0.39	0.44	0.42	0.41	0.46	0.44
Sentence	ChatGPT_PhenoB	0.61	**0.49**	**0.54**	**0.62**	**0.51**	**0.56**
Sentence	ChatGPT_HPOapi	0.55	0.48	0.51	0.54	0.50	0.52
Report	ChatGPT_raw	0.53	0.34	0.42	0.46	0.40	0.43
Report	ChatGPT_PhenoB	**0.75**	0.34	0.47	**0.62**	0.40	0.48
Report	ChatGPT_HPOapi	0.66	0.34	0.45	0.54	0.40	0.46

TABLE II: Experiment #1 - ID-68 Results

	System	Micro-Average			Macro-Average		
		P	R	F1	P	R	F1
	Clinphen	0.64	0.41	0.50	0.51	0.41	0.45
	MetaMap Lite	0.69	0.48	0.57	0.63	0.49	0.55
	PhenoTagger	0.79	0.63	0.70	0.78	0.68	0.73
	PhenoBERT	**0.80**	**0.67**	**0.73**	**0.79**	**0.71**	**0.75**
Sentence	ChatGPT_raw	0.26	0.23	0.24	0.22	0.22	0.22
Sentence	ChatGPT_PhenoB	**0.60**	**0.33**	**0.43**	**0.53**	**0.33**	**0.41**
Sentence	ChatGPT_HPOapi	0.48	0.29	0.36	0.40	0.28	0.33
Report	ChatGPT_raw	0.29	0.11	0.16	0.13	0.12	0.12
Report	ChatGPT_PhenoB	0.54	0.13	0.21	0.24	0.14	0.18
Report	ChatGPT_HPOapi	0.47	0.12	0.19	0.21	0.13	0.16

TABLE III: Experiment #1 - GSC+ Results

However in all cases, ChatGPT underperforms compared to the current state-of-the-art systems. The experiments #2

[4]https://github.com/EclipseCN/PhenoBERT/blob/main/phenobert/utils/
[5]https://clinicaltables.nlm.nih.gov/apidoc/hpo/v3/doc.html

and #3 (Fig. 3) allow to derive some hypothesis explaining this performance. Note that only micro-average results are shown as the tendencies are exactly the same for macro-average.

As expected, changing the prompt modifies the final results. Prompt#2 improves all metrics, both at report and sentence levels, but has a higher influence at sentence level. We may emphasize that prompt is less predominant when the input text is large, as a wider context limits the intent interpretation. Although it is difficult to bring a clear conclusion, these results show the importance of the prompt engineering part, which is inherent to this type of model.

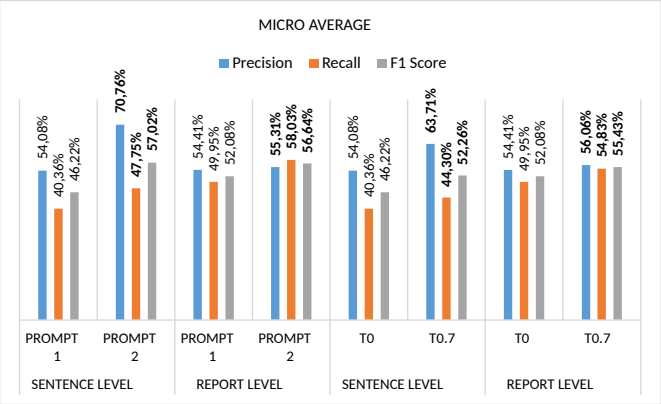

Fig. 3: Prompt and Temperature Influence

In addition, increasing temperature to 0.7 improves slightly the scores, with a similarly predominant influence on sentence level approach. When temperature is low, the model is more deterministic, choosing systematically the most probable string matching parts of the input text, which works well when phenotypes are explicitly written. When the reference is more implicit, the learned distribution does not fit well, and randomness can improve the matching from time to time.

It is worth reminding here that the evaluation method does not take into account the ontology distance between a candidate and the target phenotype. The use of a weighted generalized match [6] might increase the score, and reduce the differences with other systems.

Another limitation might be the number of maximum tokens set in the query: it could be interesting to test the impact of several values for this parameter.

Finally, the very nature of the dataset can have a significant influence. It is especially the case for the two target dataset in experiment #1: ChatGPT perfoms much better on ID-68 than on GSC+.

V. CONCLUSIONS

ChatGPT has proven its ability to tackle many tasks in a wide variety of domains. However, its performance on phenotypes extraction from text based on a reference ontology does not meet expectations, as it underperforms quite clearly compared to dedicated systems. Even if it remains difficult to determine all causes explaining these results, this work highlighted several limitations.

First of all, ChatGPT lacks factual knowledge such as HPO structure and relationships, and the association between an extracted phenotype string and its corresponding ID happens to be spurious. This is expected since the model learned a statistical distribution of tokens, and does not encode the facts carried by knowledge sources.

Moreover, such a model is very versatile: modifying prompt and temperature impacts the scores, sometimes significantly when dealing with a sentence level approach. It is therefore difficult to choose the best set of parameters, except by performing many configuration experiments, which can be tedious especially with prompts values. Prompt design should be considered a crucial step for IE with ChatGPT.

Hence, ChatGPT seems not to be relevant for the considered task yet, as confirmed by the presented experiments on two different datasets. We believe a possible improvement could be to incorporate factual knowledge and referenced sources, that may counterbalance the highlighted limitations.

While waiting for these potential improvements, this study confirms the value of dedicated approaches rather than over-generalized models to achieve good performances in NLP applications such as IE.

REFERENCES

[1] Anazi Shams & all. Expanding the genetic heterogeneity of intellectual disability. *Human genetics*, 136:1419–1429, 2017.

[2] Arbabi Aryan & all. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR medical informatics*, 7(2):e12596, 2019.

[3] Bojanowski Piotr & all. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[4] Brown & all. Language models are few-shot learners, 2020.

[5] Clement Jonquet & all. The open biomedical annotator. *Summit on Translational Bioinformatics*, 2009:56 – 60, 2009.

[6] Cong Liu & all. Ensembles of natural language processing systems for portable phenotyping solutions. *Journal of Biomedical Informatics*, 100:103318, 2019.

[7] Habibi Maryam & all. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[8] Jacob Devlin & all. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] Jinhyuk Lee & all. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, sep 2019.

[10] Jung Hoon Son & all. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. *American Journal of Human Genetics*, 103:58–73, 7 2018.

[11] Köhler & all. The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, 12 2020.

[12] Ling Luo & all. Phenotagger: A hybrid method for phenotype concept recognition using human phenotype ontology. *CoRR*, abs/2009.08478, 2020.

[13] Manuel Lobo & all. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017:1–8, 2017.

[14] Ouyang & all. Training language models to follow instructions with human feedback, 2022.

[15] Vani Ankit & all. Grounded recurrent neural networks, 2017.

[16] Y. Feng & all. Phenobert: a combined deep learning method for automated recognition of human phenotype ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, (01):1–1, apr 5555.

[17] Yuan & all. Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. *Briefings in Bioinformatics*, 23(2), 02 2022. bbac019.

[18] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.