

Enhancing Phenotype Recognition in Clinical Notes Using Large Language Models: PhenoBCBERT and PhenoGPT

Jingye Yang^{1,5}, Cong Liu², Wendy Deng¹, Da Wu¹, Chunhua Weng²,
Yunyun Zhou^{1,3,*}, Kai Wang^{1,4,**}

¹ Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

² Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA

³ Biostatistics and Bioinformatics facility, Fox Chase Cancer Center, Philadelphia, PA 19111, USA

⁴ Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

⁵ Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104, USA

*: Corresponding author. Email: yunyun.zhou@fccc.edu

** : Corresponding author and lead contact. Email: wangk@chop.edu

SUMMARY

To enhance phenotype recognition in clinical notes of genetic diseases, we developed two models - PhenoBCBERT and PhenoGPT - for expanding the vocabularies of Human Phenotype Ontology (HPO) terms. While HPO offers a standardized vocabulary for phenotypes, existing tools often fail to capture the full scope of phenotypes, due to limitations from traditional heuristic or rule-based approaches. Our models leverage large language models (LLMs) to automate the detection of phenotype terms, including those not in the current HPO. We compared these models to PhenoTagger, another HPO recognition tool, and found that our models identify a wider range of phenotype concepts, including previously uncharacterized ones. Our models also showed strong performance in case studies on biomedical literature. We evaluated the strengths and weaknesses of BERT-based and GPT-based models in aspects such as architecture and accuracy. Overall, our models enhance automated phenotype detection from clinical texts, improving downstream analyses on human diseases.

Keywords:

Human Phenotype Ontology, named entity recognition, transformer, clinical notes, electronic health records

INTRODUCTION

Rare diseases affect 30 million people in the USA and more than 300–400 million worldwide, often causing chronic illness, disability, and premature death¹⁻³. Phenotype-driven approaches are increasingly used to facilitate the genetic diagnosis of rare diseases^{1,4,5}. For example, a number of computational methods have been developed to facilitate phenotype-based prioritization of disease variants and genes⁶⁻¹², and some methods also enable the prediction of Mendelian diseases directly from phenotype information¹³⁻¹⁵. To facilitate computational phenotype analysis, the Human Phenotype Ontology (HPO) was established, which provides a standardized vocabulary to describe phenotypic abnormalities in human diseases¹⁶. The current release of HPO (June 2022) covers 13,000 terms and over 156,000 annotations on hereditary diseases. However, the curation of the HPO largely relies on experts' input, which may not be comprehensive enough to capture all the clinical terms, such as the age-specific neurodevelopmental and neuropsychiatric phenotypes. For instance, our previous study on children with autism spectrum disorders (ASD)¹⁷ developed a natural language processing (NLP) pipeline and identified more ASD terms than those documented in the HPO (i.e., 3,000 terms linked to ~2,000 unique Unified Medical Language System (UMLS) concepts), representing one of the largest ASD terminology available to date. This is an illustration that NLP techniques can facilitate human reviewers in building catalogs of phenotype terms that are not previously documented for specific diseases. Therefore, there is a strong need to develop more efficient NLP methods to extract novel phenotypic concepts that may not be well covered in HPO.

Automated phenotype concept recognition from unstructured biomedical text is a type of named entity recognition (NER) task and remains a challenging problem in the biomedical NLP field. As shown in **Table 1**, we summarized three major types of methods to tackle this problem: (1) rule-based (string matching, dictionary-based, statistical model, etc.) algorithms, (2) machine learning algorithms, including recently developed deep learning methods, and (3) hybrid models combining both approaches. The first generation of tools for clinical concept recognition were either dictionary-based or rule-based approaches, such as MetaMap¹⁸, NCBO annotator¹⁹, ClinPhen²⁰, and the Aho-Corasick algorithm used in Doc2HPO²¹. MetaMap is a program developed at the National Library of Medicine (NLM) to map biomedical texts to the Metathesaurus based on a knowledge intensive approach. NCBO Annotator first creates direct annotations from raw text based on syntactic concept recognition according to a dictionary that uses terms (concept names and synonyms) from both UMLS and NCBO BioPortal ontologies, and then different components expand the first set of annotations using the knowledge represented in one or more ontologies. ClinPhen uses sequential analytic procedures with a rule-based NLP system to decide which phenotypes correspond to true mentions and which are false positives. CLAMP (Clinical Language Annotation, Modeling, and Processing) is a clinical NLP toolkit that provides not only state-of-the-art NLP components, but also a user-friendly graphic user interface that can help users quickly build customized NLP pipelines including phenotype recognition tasks²². In recent years, researchers began to adopt machine-learning models, including deep-learning models, due to their high accuracy and independence of hand-crafted features²³. Machine learning based approaches, such as conditional random field (CRF)²⁴ and support vector machine (SVM)²⁵, were developed to advance the novel concept discovery; but more recently, convolutional neural network (CNN), recurrent neural networks (RNN)²⁶, and transformer are increasingly used. For example, the neural concept recognizer (NCR) uses the CNN to encode input phrases and then rank medical concepts based on the similarity in that vector space²⁷. Transformer allows for parallelization and makes it possible to train on a much larger corpus and recognizes the long-range relationship in a sentence via its

attention mechanism²⁸. Transformer-based models are particularly useful for phenotype recognition since the contextual information can be very useful in determining whether a set of phrases represents a description of a particular phenotype. Pre-trained language model BERT (Bidirectional Encoder Representations from Transformers)²⁹ is the most well-known transformer-based model, using general corpora as the training set. The BERT_{BASE} model has 12 encoders with 12 bidirectional self-attention heads, and the BERT_{LARGE} model has 24 encoders with 16 bidirectional self-attention heads. Both models are pre-trained from unlabeled data extracted from the BooksCorpus with 800M words and English Wikipedia with 2,500M words.

Several studies have explored the utility of BERT-based approaches in the clinical and biomedical domains for the concept recognition task. In the biomedical domain, BioBERT³⁰ continued to train a model on ~18 billion words from PubMed research papers using BERT pre-trained model as the base model. BioBERT outperforms BERT in a variety of biomedical text mining tasks, which suggests that continued training on domain specific corpora improves the performance. To make the BERT model more useful in the clinical field, ClinicalBERT³¹ (updated now as Bio+Clinical BERT on Huggingface model cards) started from BioBERT's pre-trained model and continued to train on 3 million notes from the MIMIC III corpora³². Similarly, BERN³³ used BioBERT as the backbone model to recognize known, novel, and multiple types of biomedical entities, such as gene, disease, drugs; later, the same team developed an improved version BERN2³⁴ for NER using bio-lm pre-trained model. Note that these two methods focused on fine-tuning the BERT-based pre-train model for the discovery of multiple types of biomedical entities. PhenoBERT is a combined deep learning method for automated recognition of HPO; it introduces a two-levels CNN module consisting of a series of CNN models organized in two levels³⁵. Furthermore, PhenoTagger was developed for HPO recognition using both dictionary-based and BERT-based model³⁶, and is currently one of the most reliable methods for automated HPO extraction from biomedical texts. We note that PhenoTagger can be improved in several aspects. First, PhenoTagger relies on n-gram phrase classification, so it can be difficult to differentiate different (context-dependent) concepts with the same texts or the same concept with different expressions; second, it is difficult to recognize the misspelling or low-frequency phenotypes in the samples; third, negation is not supported yet the negated instances can be important in diagnosing diseases.

As a transformer encoder-based model, BERT has shown excellent performance on a wide range of tasks in biomedical NLP applications, including named entity recognition (described above), text classification (such as disease prediction³⁷), and relation extraction (such as chemical-protein relation extraction³⁸). However, little work has been done to evaluate transformer decoder-based models, e.g., GPT-type models, in biomedical NLP tasks. In our view, this is partly due to the following reasons:

1. The BERT structure was pre-trained on a large corpus so that it can be fine-tuned for a wide range of specific NLP tasks. In contrast, GPT was initially designed primarily for generating text such as in chatbots that fits the desired output format.
2. BERT was released by Google in 2018 and quickly became widely available to the research community. Although GPT was initially released by OpenAI in 2018, it is less well-known and it took a few years to improve until ChatGPT becomes widely successful. Moreover, GPT-based models were initially only available for research purposes, with limited access to commercial users.

3. Recent versions of GPT models are typically much larger and require significantly more computational resources compared to BERT models, which make them less accessible for researchers and practitioners with limited resources and budgets.

Nevertheless, recent advances in the ability to scale Large Language Models (LLMs) have resulted in top-notch performance across various NLP tasks³⁹. The ability to scale LLMs to hundreds of billions of parameters has unlocked additional capabilities such as in-context few-shot learning, making it possible for LLMs to perform well on tasks trained on only a handful of examples⁴⁰. Chain-of-Thought (CoT) prompting has also showcased the robust reasoning ability of LLMs across a wide variety of tasks, even in the absence of few-shot examples⁴¹. Additionally, Huang et al. have demonstrated that LLMs are capable of self-improving with only unlabeled datasets⁴².

Current phenotype recognition models primarily rely on the HPO dictionary for training or prediction, limiting their ability to capture all the important phenotypic features, especially for those not well represented by HPO terms. Therefore, it will be of great scientific and practical importance to develop an accurate phenotype concept recognition model to extract all available phenotype information from clinical notes. In this study, we proposed two transformer-based models for phenotype concept recognitions: PhenoBCBERT (BERT-based) and PhenoGPT (GPT-based), and compared their performance to other existing models. We initially trained a phenotype recognition model on top of Bio+ClinicalBERT for rare disease-specific NLP text mining tasks, allowing the recognition of terms outside of the standard HPO vocabulary. Next, we implemented several GPT-based phenotype recognition models to supplement BERT-based models. Compared to existing tools such as PhenoTagger, our PhenoBCBERT can accurately infer essential phenotypic features from given contexts, despite the occurrence of non-HPO phenotypes, misspellings, and lexical dissimilarities with the original training data. Meanwhile, PhenoGPT can achieve comparable results with PhenoBCBERT with significantly fewer fine-tuning data. Therefore, both of our Transformer-based models are robust and can complement each other to achieve improved performance in concept recognition of HPO or non-HPO phenotypes.

RESULTS

Summary

In the sections below, we first separately described our evaluation of the performance of BERT-based (PhenoBCBERT) and GPT-based (PhenoGPT) models with comparisons to existing approaches. We then performed a comparative analysis of BERT-based and GPT-based models and discussed the relative merits of each model's architecture. Finally, we demonstrated real-world applications of these methods on publicly available clinical notes in the published genetics literature. The workflow design of the project is shown in [Figure 1](#). The model architecture of PhenoBCBERT and PhenoGPT are illustrated in [Figure 2](#). The tokenization techniques used in two models are described in [Table 2](#).

Evaluation of PhenoBCBERT

For the evaluation of PhenoBCBERT, our initial dataset contains 3400 automatically labeled clinical notes and 460 hand-labeled clinical notes exclusively from the CHOP database, among

which 200 automatically labeled ones were randomly selected for testing and the remaining ones were used for training.

Comparing PhenoBCBERT and PhenoTagger

We have on average ~79% overlap of concepts identified from our PhenoBCBERT model and PhenoTagger, among the 200 test notes. By examining the test notes where >50% of entities were found in PhenoTagger but not found in PhenoBCBERT, we inferred the following two scenarios with specific examples:

1. PhenoTagger's results may contain repeated/nested phenotype mentions, yet they are skipped by our model. For instance, as shown in the clinical note 4 of **Figure 3**, PhenoTagger infers both "autism" and "autism spectrum disorder" while our PhenoBCBERT only infers "autism spectrum disorder".
2. PhenoTagger's results may contain non-phenotype entities (false positives) from human review. For instance, as shown in the clinical note 2 of **Figure 3**, PhenoTagger includes "contact the Roberts", while PhenoBCBERT does not.

There are several possibilities to explain the observed results. PhenoTagger, like other rule-based or hybrid model^{35,43}, utilizes localized information to facilitate predictions. Specifically, it will segment a sentence into small groups of n-grams ($n = 2\sim 10$), and then feed into a deep-learning model (e.g., BioBERT) for classification. This output will be combined with dictionary-based inference for the final prediction. Since segments of local information and dictionary-based matching rely on the restricted meaning and patterns of short segments of tokens/words, it renders this method sensitive to the similarity of strings between input data and phenotype entities, which in turn leads to over-matching of the same tokens regardless of its meaning and its occurrence.

In contrast, our PhenoBCBERT takes the complete long-stretch of meaningful sentences or paragraphs as the input, passes into a fine-tuned deep-learning model, and then post-processes its output with minimal intervention (resemble subwords, etc.). Hence it will make a positive prediction only if the surrounding tokens grant it semantic meanings that resemble phenotype mentioning. Meanwhile, our infused hand-labeled data is continuously correcting the model from aligning to any wrong classifications automatically generated by PhenoTagger.

We further examined the cases where entities were identified in PhenoBCBERT but not found in PhenoTagger. We illustrated a few case studies in **Figure 3** as examples. In most cases, entities either do not have strong lexical similarity to a standard HPO term, or are written in a non-standard format (abbreviations, uncommon synonym, misspellings, etc.). For instance, in the clinical note 7 of **Figure 3**, our PhenoBCBERT successfully detected "difficulty urinating", which corresponds to "urinary retention" in the standard HPO format (HP:0000016).

PhenoTagger uses the standard HPO dictionary as the training data, which includes limited amounts of standardized phenotypic abnormalities encountered in human diseases. Therefore, it is not surprising that our model can reveal many phenotype entities that are not documented in any standard dictionaries. We acknowledge that our model may also generate false positives, such as 'fragile' in Note 3. With the help of this pipeline, we will match newly found phenotype entities with higher-level HPO terms and it may help expand phenotype catalogs of rare diseases.

In addition, we found that PhenoBCBERT is more robust to typos and misspelling of words in the sentence. After removing, or replacing, random letters and/or words of phenotype entities in the testing sentence, PhenoBCBERT still recognize the compromised phenotype entities. For example, by replacing "palmar telangiectasia" with "pal telrngiectasia", PhenoTagger cannot recognize this phenotype while our model will recognize it precisely. In addition, our model can

recognize abbreviations or short notation by doctors, which are not standard HPO terms. For example, 'severe IUGR' (Note 5) is not recognized by PhenoTagger.

Evaluation of PhenoGPT

As shown in [Table 3 and 4](#), even though PhenoGPT used a significantly smaller volume of a public dataset in comparison to the in-house training data used by the PhenoBCBERT model, PhenoGPT has demonstrated highly competitive results, with an impressive accuracy of 0.857 and the top F1 score (GPT-J based) when evaluated on the BiolarkGSC+ validation dataset. In the context of prompt-based learning, we have implemented a one-shot learning strategy and through this approach, we found that the overall size of the model plays a critical role in determining its effectiveness and overall performance. GPT-J is the smallest model with 6 billion parameters; GPT-3 (davinci) is the most capable GPT-3 model with 175 billion parameters⁴⁰; GPT-3.5 (gpt-3.5-turbo) is one of the most advanced GPT model built on GPT-3 and involves reinforcement learning with human feedback (RLHF)⁴⁴, optimized for chat. The outputs generated by GPT-J would produce incorrect phenotype entities along with arbitrary HPO IDs. In comparison, the GPT-3.5 model is capable of generating more than 80% of accurate phenotype entities, albeit with slightly compromised HPO IDs.

After the prompt-based learning, three models have drastically different performance as shown in the “GPT comparison 1” panel of [Figure 4](#). The GPT-J has very low recall and makes up fake HPO IDs, whereas both GPT-3 and GPT-3.5 have reasonably good performance on entity extraction. On the other hand, GPT-3.5 is capable of performing entity normalizations by accurately assigning the appropriate HPO ID to corresponding entities. For instance, it can correctly associate "talipes equinovarus" with HP_0001762 and "foot deformities" with HP_0001760, likely due to its exposure to public phenotypic datasets during the pretraining process. Conversely, GPT-3 appears incapable of achieving such entity normalization without undergoing the fine-tuning process. However, after fine-tuning, GPT-3 is capable of successfully performing entity normalization, as discussed below.

The prediction results after fine-tuning are illustrated in the GPT comparison 2-6 of [Figure 4](#). We found that the performance of the model is significantly influenced by the size of the fine-tuning dataset. Yet, both the closed-source and open-source models can yield comparably favorable outcomes. As shown in the “GPT comparison 2” panel ([Figure 4](#)), the performance of the GPT-3 based model fine-tuned on 28 instances is even worse than the prompt-based learning. Conversely, other model finetuned on 200 instances has considerably better results. It has fewer false positive predictions and more accurately normalized HPO IDs compared to prompt-based results. For example, the fine-tuned GPT-3 (w/ 200 instances) can successfully extract “distal arthrogryposis (HP: 0005684)” and “foot deformities (HP: 0001760)”, which were in the results of prompt-based GPT-3 prediction but with incorrect HPO IDs (HP: 0001812 and HP: 0000827, respectively); the fine-tuned model also excluded false outcomes in the prompt-based learning like “strict diagnostic criteria”, “incomplete ascertainment”, “range of phenotypes”, etc. It is not surprising that the fine-tuned model will be more reliable and accurate on the phenotypic entity recognition than prompt-based learning, since the model will compute gradients and make updates on all of its parameters to fit the fine-tuning dataset. In addition, we searched through the 200 notes in the fine-tuning dataset and found that all phenotypic entities with the correct HPO ID normalization have appeared at least once in the fine-tuning dataset, which explains the reason why the fine-tuned model can successfully normalize those phenotypes.

As illustrated in the GPT “comparison 3” panel ([Figure 4](#)), both the closed-source and open-source models demonstrate strong ability to identify phenotype entities accurately. For example,

in the case of the previous 9 positive phenotypic entities, both the GPT-J-based and Falcon-based open-source models were able to successfully identify all of them. The LLaMA-based open-source model, however, missed one phenotype, "arthrogryposis", but it still correctly classified it under the more comprehensive HPO term, distal arthrogryposis (HP: 0005684). It is important to note that both LLaMA and GPT-J generated an additional false positive each; LLaMA misidentified "incomplete ascertainment", and GPT-J misidentified "variable". On the other hand, the Falcon model demonstrated an impeccable match for all phenotype entities.

We further evaluated the performance of our open-source models across various clinical abstracts, as showcased in GPT comparisons 4-6 ([Figure 4](#)). We found that these three PhenoGPT open-source models demonstrate consistency and strong performance. In rare instances, there are discrepancies in results, as seen in GPT comparison 6. The GPT-J-based model predicted two additional true positives, but also reported a false positive, "miscarriage". Both LLaMA and Falcon predicted one additional true positive each, namely, "coxa-epiphysiolysis" and "obesity".

In conclusion, the results presented above emphasize the critical role played by both the size and quality of the pretraining data in determining the final performance of the model. Furthermore, it is highly recommended to initially fine-tune the large base GPT model on domain-specific datasets before utilizing it for inference, as this process enhances the model's ability to perform accurate entity normalization. When it comes to open-source versus closed-source models, the choice largely depends on an individual's specific requirements. Open-source models offer greater flexibility, allowing for customization and the use of various training strategies such as LoRA, quantization, and others. On the other hand, closed-source models depend on third-party services and are typically easier to implement through API calls. Both types of models demonstrate commendable performance.

Comparison between the BERT-based model and GPT-based model

Performance on phenotype entity recognition

Both PhenoGPT and PhenoBCBERT are capable of extracting phenotypic information from unstructured raw clinical data. They have comparably high precision, recall and accuracy on the public validation dataset. As shown in the [Table 3 and 4](#), PhenoGPT has the best Recall and F1 score for certain datasets while PhenoBCBERT has relatively lower scores. Since we formulated NER as a causal language model task for GPT model's fine-tuning, it requires fewer efforts to post-process its outcomes (e.g., HPO ID normalization). In contrast, PhenoBCBERT needs extra modules to normalize its phenotype entity predictions (e.g., Sent2Vec or SVM). The precision scores of our models are not superior, which is partially due to the fact that our models can detect the words that are not included in the standard HPO dictionary.

We also observed the nuances in the performance of the PhenoGPT model family, which showed slightly inferior results on the ID-68 dataset compared to PhenoBCBERT. These nuances can be attributed to various factors including dataset characteristics, architectural decisions, and training approaches. Specifically, ID-68 dataset comprises real-world clinical notes focused exclusively on families with intellectual disabilities, which naturally limits the model's performance to specific areas. PhenoBCBERT excels particularly in recognizing entities related to intellectual disabilities, likely owing to the high representation of this specific phenotype in the in-house dataset. This contrasts with other larger language models, which demonstrate better overall performance, as illustrated in [Table 3](#). Another point to consider is the inherent difficulty in training large GPT-based language models. These models often rely heavily on their pre-existing knowledge base for decision-making, which is precisely why such large-scale models are developed in the first place. One of our primary objectives is to fine-tune large language models in a cost-effective manner to excel at general Named Entity Recognition

(NER) tasks in the biomedical domain. Our findings indicate that both a fully fine-tuned BERT model and an efficiently fine-tuned large GPT model can outperform other general-purpose models in these tasks. This validates our approach and bolsters our confidence to continue refining our large language models for larger cohorts in future projects.

Data efficiency for model training

Despite the differences in model architecture and training approaches between PhenoBCBERT and PhenoGPT, both models demand accurately labeled data containing phenotypic information. Therefore, gathering and preprocessing raw clinical data for both models are equally challenging and expensive. Nonetheless, since GPT has been pretrained on a substantially larger dataset compared to BERT, it needs a relatively smaller fine-tuning dataset to attain comparable outcomes. Note that it is also possible that GPT models might have already seen the public datasets that were used in the fine-tuning step.

Computing resource requirement and accessibility

PhenoBCBERT has 110 million parameters and requires 1-2GB of RAM during training, making it suitable for fine-tuning with larger batch sizes. The closed-source PhenoGPT can have up to 175 billion parameters, demanding 600 GB of RAM, so fine-tuning is performed using OpenAI's API and stored in the cloud for inference. The open-source PhenoGPT models still necessitates a certain amount of GPU resources, ranging from 14-70GB, depending on the specific model and training strategy utilized. The open-source model was saved locally and sharable for public use.

In conclusion, BERT is more affordable and efficient, while GPT has excessive capabilities. The choice of which model to use depends on task requirements, and it is advisable to test different models since they each have their own merits.

Additional validation through case studies

We applied both BERT-based and GPT-based models (GPT-3) on clinical notes selected from papers published by the American Journal of Human Genetics (AJHG)⁴⁵⁻⁴⁷ to extract phenotype entities. As shown in **Figure 5**, we noticed some differences between the results of PhenoTagger and our models. We underlined PhenoTagger-specific outputs, PhenoBCBERT-specific outputs, and PhenoGPT-specific outputs in green, red and blue, respectively. We also highlighted negation detection and misspelled entities in red and yellow.

In the example of **Figure 5A**, we observed that PhenoBCBERT captured all the phenotype entities, including two essential phenotypes missed by PhenoTagger (“restricted cerebellar growth” and “problems with latching and swelling”) and one negated phenotype entity (“chromosomal abnormalities”, which is an “abnormal cellular phenotype”). In addition, PhenoTagger mistakenly recognized one non-phenotype entity “nuchal translucency” as “Increased nuchal translucency” (HP: 0010880). PhenoGPT had high precision and skipped the negated phenotype entity, but it could not identify “ventriculomegaly” and “small cerebellum”.

In the example **Figure 5B**, PhenoTagger could not recognize phenotypes “pyelocaliceal dilatation”, “dysmorphisms” and “micrognathia”. In contrast, PhenoBCBERT successfully highlighted all essential phenotypes, with extra granularity (“learning difficulty in expressive language” instead of “learning difficulty”). We also noticed that PhenoBCBERT did not identify non-phenotype entities such as “vacuum extraction”. For this note, PhenoGPT did not over-predict phenotype terms such as “gestational diabetes” (maternal phenotype) and “umbilical artery” (non-phenotype). Additionally, PhenoGPT identified the less obvious phenotype entity, “height and weight above the 95th centile”.

In the example **Figure 5C**, PhenoTagger could recognize most of the positive phenotype entities, while both PhenoBCBERT and PhenoGPT slightly outperformed with more true positive predictions (“severe neurosensory hyposacusia”) and more accurate phenotypic description (“... with stiffness”, “right-convex scoliosis”, “Achilles tendon”, “tendon retraction”). Our models missed one general phenotype “Pyramidal signs”, but successfully captured its child term “high tendon reflexes”, which contains more information at its level of taxonomic hierarchy.

Interestingly, in the example **Figure 5B** and **Figure 5C**, we noticed misspellings in the original data (these mistakes are present in the originally published AJHG manuscript): “micrognatia” should be “micrognathia”, and “hyposacusia” should be “hyposacosis”, which disguised themselves from PhenoTagger yet both PhenoBCBERT and PhenoGPT could recognize them. These examples supported the robustness of our transformer models as discussed above.

The under-performance of PhenoTagger on these applications is mainly caused by two reasons, as specified in the original paper³⁶: first, PhenoTagger cannot disambiguate the different concepts with the same text name; second, PhenoTagger misses some phenotype concepts that are lexically dissimilar with the concept terms in the HPO.

In conclusion, these case studies demonstrate that context information can be important and informative to reliably make predictions on phenotype recognition tasks. PhenoBCBERT and PhenoGPT can accurately infer essential phenotypic information from the given context, despite the occurrence of rare phenotypes, misspelling, and lexical dissimilarity with the original training data.

DISCUSSION

The study used BERT-based and GPT-based models on clinical texts from pediatric patients to identify known and unknown (not documented by HPO) clinical phenotypes. We found that both PhenoBCBERT and PhenoGPT can identify new concepts with better accuracy, recall, and precision than competing models. The methods can be adapted to other biomedical domains and can identify novel entities based on context. Despite the strengths, these models have several limitations and further improvements can be made.

Effect of data quantity and quality

Both PhenoBCBERT and PhenoGPT require annotated training data. The performance levels of the fine-tuned models may vary based on the quality of data collected and the labeling strategy utilized. Additionally, there is a potential for training bias, especially if physicians from the same institutions or teams tend to document notes in similar styles. For instance, physicians often repeatedly include a patient's past clinical history in a current clinical note when drafting a new note, which may lead to a bias toward specific semantic structures or recurring tokens (when the same passage occurs multiple times in the same note). Additionally, unexpected errors by doctors (such as typos, missing words, or incorrect phenotypes) can hinder the automatic labeling process from generating accurate training data. Moreover, the notes may contain redundant yet potentially privacy-leaking information (such as a patient's address or first name), which raises serious privacy and ethical concerns in the model training⁴⁸. Although we used a Stanford-Penn MIDRC Deidentifier model for de-identifying all clinical notes before using them in the BERT training process, we are still concerned about the possibility of leaking private information from patients and therefore cannot share the model trained on in-house data. On the other hand, PhenoGPT, which can utilize both open-source and closed-source GPT models, was fine-tuned using only minimal amounts of data available to the public. Thus, we do not have

the same privacy leakage concerns that BERT-based models might pose. Despite this, thanks to their extensive pre-trained knowledge base, they still demonstrate strong performance on the evaluation datasets. In our future research, we aim to incorporate more diverse datasets sourced from various healthcare systems for additional refinement/improvements of the phenotype recognition models, with appropriate IRB protocols. We will assess the degree to which these institutional datasets enhance our models' effectiveness and generalizability. Lastly, we discussed above generic concerns applicable to all deep learning models when used on real clinical data. Several different existing strategies, including strings comparison for repeated tokens, auto-correction for misspelling⁴⁹, de-identification model⁵⁰ to remove patient information, have been explored to minimize the undesirable negative impacts of raw clinical data.

Model structure

For the BERT-based model, we used pretrained Bio+Clinical BERT as the main deep learning model, which is configured from BERT_{base}. The size of the model (number of attention layers, number of attention heads, dimensions of non-linear layer, etc.) is relatively small compared to recently published large language models. In this case, pretraining for a larger model with billions of parameters on EHR data may be necessary. Based on the promising result from the current study, we will pretrain a specialized large language model on de-identified clinical notes from our database for future downstream EHR-NLP tasks.

We also noticed the restrictions on the length of sentences for the language model (512 tokens for BERT_{base} and BERT_{large}). To overcome the bottleneck, we will further adopt the down-sampling⁵¹ or/and filtering methods⁵² on top of our model. Meanwhile, we can distribute many independent models on multiple GPUs for training such that a combination of them will take care of extra-long samples⁵³.

With the GPT-based model, its larger input window and automatic normalization are advantageous as part of the output. However, the vast number of parameters and associated costs can make it difficult for retraining. Moreover, as a generative language model, its strengths in chat completion and next-sentence generation were not fully utilized. In fact, our fine-tuning process diminished its capacity to generate coherent human-like responses. Consequently, a more effective strategy to fully leverage the model's structure is essential.

Model Selection and Implementation

Our PhenoBCBERT model, initialized from a BERT-based architecture, demands the fewest computational resources. However, its Named Entity Recognition (NER) training strategy necessitates additional steps for entity normalization and negation, potentially resulting in a longer overall processing time. On the other hand, while our GPT-based models do require a certain degree of computational power during training, their final versions are relatively resource-efficient and easy to deploy for inference and predictive tasks.

In comparison to PhenoTagger, both of our models can be conveniently transferred to a local machine for making predictions. Even for setups with limited GPU resources, CPU-based deployment remains a viable alternative. For users interested in custom training our models on their datasets, we have provided comprehensive guidelines and scripts in the code repository that detail how to fine-tune large language models with minimal computational expenditure.

Future Directions

Although both our models show promising results for phenotype entity recognition, they cannot classify all of the desired phenotype entities in the context. Concretely speaking, we list the following two potential improvements in terms of our current work.

Phenotype entity recognition

In the test data, we noticed that our model would recognize medication names and disease names as positive if they appear in a similar context as phenotypes. (e.g., he has **red eyes** vs. he has **sertraline**.) Also, despite post-processing, the output of BERT-based model may skip prepositions in phenotype entities (**learning difficulty** in the **expressive language**). In very few cases, both models missed out on rare phenotype entities.

Incorporating hard negative examples into the annotated dataset for fine-tuning could enhance phenotype entity recognition performance. Additionally, employing multi-modal networks to boost natural language understanding may be another potential approach. These strategies represent potential avenues for future exploration.

HPO normalization and negation

We observed that our model can encounter errors when associating phenotype entities with their corresponding HPO IDs. This issue arises as each HPO term may have multiple synonyms or typographical errors that were not present in the training data. Furthermore, PhenoBCBERT employs Sent2Vec to calculate cosine similarity between a given phenotype entity and all HPO terms, assigning the most similar HPO term to the phenotype entity. Meanwhile, PhenoGPT conducts the next token prediction based on a given phenotype entity. Neither of these approaches can effectively tackle normalization problems.

We hypothesize that employing vector embedding could enhance the HPO normalization process. In particular, we aim to leverage GPT's extensive knowledge to generate a database of HPO embeddings as high-dimensional vectors, where each vector is normalized to enable either Euclidean or cosine distance as similarity measures. Such an embedding should yield more accurate normalization outcomes and better interpretability. Lastly, we can enhance the detection of negated entities by incorporating a dedicated post-processing model based on the transformer or LSTM architectures as a part of the pipeline, rather than relying on general NLP packages such as *Negspacy*. This will be one of our focuses in future research.

In addition to the two aspects previously discussed, we have also fine-tuned the latest version of the LLaMA 2 model to investigate whether it offers any advancements in HPO normalization or notable improvements in phenotype entity recognition. Unfortunately, the most recent LLaMA 2 iteration does not yield better outcomes; its performance closely mirrors that of the original LLaMA. We have made this model accessible in our public code repository for further testing. Going forward, we plan to delve into the utility of large-scale generative language models for a range of biomedical NLP tasks, such as entity recognition, relationship extraction, and text classification, in a cost-effective manner. Given the strength of reinforcement learning, we also hope to develop such a model capable of adapting to changes in the environment and can continue to learn over time (for example, personalized biomedical NLP pipelines). Meanwhile, scaled models will have compromised interpretability and credential issues, and carefully addressing these problems is another direction for future work.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for data should be directed to and will be fulfilled by the lead contact, Dr. Kai Wang (wangk@chop.edu).

Materials availability

This study did not generate new unique materials.

Data and code availability

The various GPT-based models and example notebook scripts are available to download at <https://github.com/WGLab/PhenoGPT>. All original code has been deposited at [DOI: 10.5281/zenodo.8346470](https://doi.org/10.5281/zenodo.8346470) and is publicly available as of the date of publication. Any additional information required to reanalyze the publicly available data reported in this paper is available from the lead contact upon request.

Summary

PhenoBCBERT utilized encoder-based BERT model by fine-tuning the pretrained Bio+Clinical BERT model for concept recognition. PhenoGPT utilized decoder-based GPT model (GPT-J, Falcon⁵⁴, LLaMA⁵⁵, GPT-3⁴⁰, GPT-3.5⁴⁴) and can be trained by either prompt-based strategy or fine-tuning strategy. The framework of the project design is shown in **Figure 1**. The workflow of both PhenoBCBERT and PhenoGPT is illustrated in **Figure 2**. All the computational models were built using the PyTorch module and HuggingFace⁵⁶ interface.

Data preparation

This study was approved by the institutional review board of the Children Hospital of Philadelphia (CHOP). Our in-house dataset consists of clinical notes of rare disease patients seen at CHOP. We used 2252 ICD10 code⁵⁷ (that can map to the rare disease in the Orphanet Database) to query rare disease patients. Among these, we sub-sampled more than 4500 clinical notes regarding rare diseases, each of which contains potential phenotypic concepts that have been recognized by physicians. Although these potential phenotype entities are not complete, they will serve as the foundational phenotype information to assist in locating clinical abstracts and incorporating additional related phenotype entities using the mixed labeling strategy outlined in the following section. Since patients might have multiple visits to the hospital due to different reasons, we only retained disease progression summary notes that map to these rare-disease specific ICD10 code. We further excluded low-quality notes defined as patients with less than 4 visits or notes with less than 1000 words. Each progression summary note contains full counseling history, including medical history, surgical history, testing and imaging, developmental history, physical history, family history, genetic counseling summary, etc. In the end, we obtained 3860 high-quality clinical abstracts by using associated fundamental phenotypic concepts as the query to extract the counseling summary, removing extraneous information, and then truncating each summary to a maximum of 2400 characters (approximately 500 tokens). This was necessary as we are utilizing a BERT-based model that can only accommodate up to 512 tokens per sample. After truncation, the 3860 clinical abstracts contained over 14,000 individual trainable sentences. Due to confidentiality concerns, we used this data locally only for PhenoBCBERT.

Our publicly available dataset consists of BiolarkGSC+⁵⁸ and ID-68³⁵. BiolarkGSC+ is an updated version of Bio-Lark gold-standard corpus (GSC) dataset⁵⁹ with 228 de-identified clinical notes abstracts and the corresponding HPO terms. The ID-68 dataset includes 68 de-identified clinical notes from families with intellectual disabilities⁶⁰ and with HPO terms annotated in the same way as in the GSC+ dataset³⁵. These public datasets were used mainly

for fine-tuning, validating GPT-based models, and comparison between various rule-based and deep-learning models. Furthermore, in our case studies, we also evaluated performance of different methods on three distinct clinical abstracts of published biomedical literature in the American Journal of Human Genetics (AJHG)⁴⁵⁻⁴⁷ as the independent third-party data source. Specifics will be covered in the extra case study section.

Tokenization

In the PhenoBCBERT, we used WordPiece tokenization and three different embeddings, consisting of token, position, and sentence embedding, to represent the input information²⁹. Most meaningful words are kept and the other words are tokenized into pieces. An uncommon word can be split into more than one tokens, and two sharps (##) are added in front of the tokens. For example, “arthritis” is tokenized into “art”, “##hr”, “##itis”. This means the word “arthritis” was less common than other words when training the WordPiece representation. This technique helps us to tokenize all possible words in the literature regardless of whether they occurred before. Two special tokens [CLS] and [SEP] were used to mark the start and the end of a sentence. We did not consider the order of sentences in our model, so the sentence embedding will default to all 0’s.

In the PhenoGPT, we used GPT’s fast Byte Pair Encoding (BPE), which can handle out-of-vocabulary (OOV) words in a similar fashion as WordPiece tokenization with additional SentencePiece⁶¹ mechanism. Since GPT is a generative decoder model, we do not need sentence embeddings. **Table 2** summarized the tokenization techniques used in two models.

Labeling and Training strategy

To train PhenoBCBERT and PhenoGPT models for phenotype entity recognition, different labeling strategies were required due to the fundamental differences in their model structures.

PhenoBCBERT

We labeled our BERT-based model PhenoBCBERT following the standard NER (name entity recognition) task practice. We utilized PhenoTagger automatic labeling in our training process because (1) it is a hybrid model combining both deep-learning-based results and rule-based classification, therefore making it a good representative of two major classification methods; and (2) it outperforms other state-of-the-art methods for phenotype concept recognitions, like Doc2hpo²¹, OBO⁶² and NCR²⁷.

To avoid duplicating PhenoTagger’s results, we will use data augmentation, as in various NLP tasks⁶³⁻⁶⁶, to improve our training dataset by infusing the 3400 automatically labeled clinical notes with 460 hand-labeled clinical notes from our in-house dataset. We name this labeling strategy with data augmentation as mixed-supervision, in contrast to a fully-supervised dataset with manually labeled clinical notes or a dataset with labels generated completely by third party NER machines.

Finally, we located starting and ending positions of each phenotype entity from PhenoTagger’s output file in the original input sentence and matched them to the tokenized subwords for labeling as shown in the table. Instead of the IOB labeling strategy which distinguishes beginning and inside tokens, we simply adopted a binary inside-outside labeling strategy for efficiency and labeled all positive sub-word tokens as 1. The final labeling is illustrated in **Table 2**. The starting and ending position are manually annotated in 460 hand-labeled clinical notes.

PhenoGPT

Instead of training a GPT model for name entity recognition, we labeled the training data to comply with the model's nature as a generative decoder. For a given clinical abstract, we generated a text by appending phenotype entities with their associated HPO IDs to the abstract for either prompt-based learning or fine-tuning.

Comparing these two models' training strategies, although NER needs a lot of time and effort on data preprocessing (mixed labeling) and labeling, it is much easier to train than causal language modeling, as in a GPT model. Causal language modeling cannot attend to future tokens and can only make predictions on the next token, whereas the NER model can attend to a complete sentence to make a classification on the [CLS] token. In general, generative language models require a larger amount of pretraining data due to a larger amount of training parameters, which may cause unexpected behaviors⁶⁷.

Model initialization, fine-tuning and prompt-based learning

Our PhenoBCBERT was initialized from the Bio+Clinical BERT³¹ as the pre-trained base model. After initialization, we continued to finetune all parameters with our mixed-supervised dataset and apply an extra token classification layer for label prediction (out-of-bag or phenotype entity). Lastly, we compared our results with PhenoTagger's output to map common phenotypic entities and rely on Sent2Vec's similarity score⁶⁸ of embeddings to normalize newly detected phenotypes in a manner similar to this study¹⁷. We also applied Negspacy⁶⁹ to exclude negated entities in the final step.

Our PhenoGPT model was constructed upon a range of GPT models, with the specific choices influenced by factors such as model availability and size, as we aimed to ensure ease of reproducibility for users. For the open-source models, we utilized GPT-J-6B, Falcon-7B, and LLaMA-7B as the initial models. In contrast, for the closed-source models, we opted to start with the GPT-3 model. Both versions were subsequently fine-tuned using the public BiolarkGSC+ dataset. It is also worth mentioning that we did not employ any proprietary data in this task to avoid the potential risk of disclosing confidential information, particularly with the closed-source model. We used cross-entropy loss to penalize the causal language model for producing incorrect next tokens, i.e., wrong phenotype entities or HPO IDs. Since a large language model like GPT is capable of prompt-based few-shot learning, we also tested GPT-J, GPT-3, and GPT-3.5 for their performance given only prompts. We used the prompt to guide the GPT model towards generating phenotypic features. For example, the prompt we used is "*please identify human phenotype ontology for me*". To evaluate the performance, we used the publicly available BiolarkGSC+ dataset and ID-68 dataset for training and validation, and both datasets include human labeled HPO terms and their identifiers. We divided the BiolarkGSC+ dataset into two asymmetric portions, one containing 200 instances and the other one containing 28 instances. We then proceeded to fine-tune our entire suite of PhenoGPT models on a larger dataset comprising 200 instances. Furthermore, to assess the impact of training sample sizes, we also fine-tuned two PhenoGPT models based on GPT-3: one was trained using a dataset of 200 instances and validated on a set of 28 instances over 12 epochs, while the other was trained on a smaller set of 28 instances and validated on the larger 200-instance set over 32 epochs. Note that these two models were also referred to as GPT-3 (200) and GPT-3 (28) in the **Figure 4**.

In the deployment of open-source PhenoGPT models (GPT-J, Falcon, LLaMA), we used 4-bits and 8-bits⁷⁰ model quantization strategy to reduce our model size to only 1/4th without compromising the performance. Following this, the language model parameters are set as fixed and a comparatively small quantity of trainable parameters are incorporated into the model via Low-Rank Adapters (LoRA)⁷¹. During the fine-tuning process, the optimizer propagates

gradients through the static 4-bit quantized pre-trained language model into these Low-Rank Adapters. It is important to note that only the LoRA layers undergo updates during the training phase. This training strategy, known as QLoRA⁷², is implemented across all of our open-source PhenoGPT models.

ACKNOWLEDGEMENTS

We thank Dr. Li Fang for providing insightful suggestions and guidance. This study is in part supported by NIH grant LM012895, HG012655, HG013031, Federal work-study program, Penn DDDI fellowship program and CHOP Research Institute. We thank technical support from the IDDRC Biostatistics and Data Science core (HD105354), and CHOP Information Services for support on GPU computing.

AUTHOR CONTRIBUTIONS

Conceptualization, J.Y., C.W., Y.Z. and K.W.; Methodology, J.Y., Y. Z., and K.W.; Software, J.Y. Investigation, J.Y., W.D., Y.Z., and K.W.; Data Curation, J.Y., W.D., and Y.Z.; Writing – Original Draft, J.Y., D.W., Y.Z., and K.W.; Writing –Review & Editing, J.Y., C.L., W.D., D.W., Y.Z. and K.W.; Resources, K.W.; Supervision, K.W.; Funding Acquisition, K.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Marwaha, S., Knowles, J.W., and Ashley, E.A. (2022). A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med* 14, 23. 10.1186/s13073-022-01026-w.
2. Groft, S.C., Posada, M., and Taruscio, D. (2021). Progress, challenges and global approaches to rare diseases. *Acta Paediatr* 110, 2711-2716. 10.1111/apa.15974.
3. Zanello, G., Chan, C.H., Pearce, D.A., and Group, I.R.W. (2022). Recommendations from the IRDiRC Working Group on methodologies to assess the impact of diagnoses and therapies on rare disease patients. *Orphanet J Rare Dis* 17, 181. 10.1186/s13023-022-02337-2.
4. Smedley, D., and Robinson, P.N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med* 7, 81. 10.1186/s13073-015-0199-2
5. Hartley, T., Lemire, G., Kernohan, K.D., Howley, H.E., Adams, D.R., and Boycott, K.M. (2020). New Diagnostic Approaches for Undiagnosed Rare Genetic

- Diseases. *Annu Rev Genomics Hum Genet* 21, 351-372. 10.1146/annurev-genom-083118-015345.
6. Yang, H., Robinson, P.N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods* 12, 841-843. 10.1038/nmeth.3484
 7. Kelly, C., Szabo, A., Pontikos, N., Arno, G., Robinson, P.N., Jacobsen, J.O.B., Smedley, D., and Cipriani, V. (2022). Phenotype-aware prioritisation of rare Mendelian disease variants. *Trends Genet* 38, 1271-1283. 10.1016/j.tig.2022.07.002.
 8. Zhao, M., Havrilla, J.M., Fang, L., Chen, Y., Peng, J., Liu, C., Wu, C., Sarmady, M., Botas, P., Isla, J., et al. (2020). Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom Bioinform* 2, lqaa032. 10.1093/nargab/lqaa032.
 9. Peng, C., Dieck, S., Schmid, A., Ahmad, A., Knaus, A., Wenzel, M., Mehnert, L., Zirn, B., Haack, T., Ossowski, S., et al. (2021). CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genom Bioinform* 3, lqab078. 10.1093/nargab/lqab078.
 10. Robinson, P.N., Köhler, S., Oellrich, A., Sanger Mouse Genetics, P., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* 24, 340-348. 10.1101/gr.160325.113.
 11. Robinson, P.N., Ravanmehr, V., Jacobsen, J.O.B., Danis, D., Zhang, X.A., Carmody, L.C., Gargano, M.A., Thaxton, C.L., Core, U.N.C.B., Karlebach, G., et al. (2020). Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *Am. J. Hum. Genet.* 107, 403-417. 10.1016/j.ajhg.2020.06.021.
 12. Birgmeier, J., Haeussler, M., Deisseroth, C.A., Steinberg, E.H., Jagadeesh, K.A., Ratner, A.J., Guturu, H., Wenger, A.M., Diekhans, M.E., Stenson, P.D., et al. (2020). AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.* 12. 10.1126/scitranslmed.aau9113.
 13. Havrilla, J.M., Liu, C., Dong, X., Weng, C., and Wang, K. (2021). PhenCards: a data resource linking human phenotype information to biomedical knowledge. *Genome Med* 13, 91. 10.1186/s13073-021-00909-8.
 14. Zhai, W., Huang, X., Shen, N., and Zhu, S. (2022). Phen2Disease: A Phenotype-driven Semantic Similarity-based Integrated Model for Disease and Gene Prioritization. *BioRxiv* doi: <https://doi.org/10.1101/2022.12.02.518845>.
 15. Kohler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dolken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* 85, 457-464. 10.1016/j.ajhg.2009.09.003
 16. Kohler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gouridine, J.P., Gargano, M., Harris, N.L., Matentzoglou, N., McMurry, J.A., et al. (2019).

- Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res* 47, D1018-D1027. 10.1093/nar/gky1105.
17. Zhao, M., Havrilla, J., Peng, J., Drye, M., Fecher, M., Guthrie, W., Tunc, B., Schultz, R., Wang, K., and Zhou, Y. (2022). Development of a phenotype ontology for autism spectrum disorder by natural language processing on electronic health records. *J Neurodev Disord* 14, 32. 10.1186/s11689-022-09442-0.
 18. Aronson, A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, 17-21.
 19. Martinez-Romero, M., Jonquet, C., O'Connor, M.J., Graybeal, J., Pazos, A., and Musen, M.A. (2017). NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. *J Biomed Semantics* 8, 21. 10.1186/s13326-017-0128-y.
 20. Deisseroth, C.A., Birgmeier, J., Bodle, E.E., Kohler, J.N., Matalon, D.R., Nazarenko, Y., Genetti, C.A., Brownstein, C.A., Schmitz-Abe, K., Schoch, K., et al. (2019). ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet Med* 21, 1585-1593. 10.1038/s41436-018-0381-1.
 21. Liu, C., Peres Kury, F.S., Li, Z., Ta, C., Wang, K., and Weng, C. (2019). Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic Acids Res* 47, W566-W570. 10.1093/nar/gkz386.
 22. Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., and Xu, H. (2018). CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 25, 331-336. 10.1093/jamia/ocx132.
 23. Cho, H., and Lee, H. (2019). Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics* 20, 735. 10.1186/s12859-019-3321-4.
 24. Lafferty, J.D., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.
 25. Boser, B.E., Subramanian, I.R., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers.
 26. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning internal representations by error propagation.
 27. Arbabi, A., Adams, D.R., Fidler, S., and Brudno, M. (2019). Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. *JMIR Med Inform* 7, e12596. 10.2196/12596.
 28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *ArXiv 1706.03762 [cs.CL]*.

29. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv abs/1810.04805*.
30. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234-1240. 10.1093/bioinformatics/btz682.
31. Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. held in Minneapolis, Minnesota, USA, June. (Association for Computational Linguistics), pp. 72-78.
32. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M.M., Moody, B., Szolovits, P., Celi, L.A., and Mark, R.G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data* 3.
33. Kim, D., Lee, J., So, C.H., Jeon, H., Jeong, M., Choi, Y., Yoon, W., Sung, M., and Kang, J. (2019). A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access* 7, 73729-73740.
34. Sung, M., Jeong, M., Choi, Y., Kim, D., Lee, J., and Kang, J. (2022). BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* 38, 4837-4839. 10.1093/bioinformatics/btac598.
35. Feng, Y., Qi, L., and Tian, W. (2022). PhenoBERT: a combined deep learning method for automated recognition of human phenotype ontology. *IEEE/ACM Trans Comput Biol Bioinform PP*. 10.1109/TCBB.2022.3170301.
36. Luo, L., Yan, S., Lai, P.T., Veltri, D., Oler, A., Xirasagar, S., Ghosh, R., Similuk, M., Robinson, P.N., and Lu, Z. (2021). PhenoTagger: A Hybrid Method for Phenotype Concept Recognition using Human Phenotype Ontology. *Bioinformatics*. 10.1093/bioinformatics/btab019.
37. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* 4, 86. 10.1038/s41746-021-00455-y.
38. Weber, L., Sanger, M., Garda, S., Barth, F., Alt, C., and Leser, U. (2022). Chemical-protein relation extraction with ensembles of carefully tuned pretrained language models. *Database (Oxford)* 2022. 10.1093/database/baac098.
39. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S.R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *EMNLP 2018*, 353.
40. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877-1901.

41. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.
42. Huang, J., Gu, S.S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. (2022). Large language models can self-improve. arXiv preprint arXiv:2210.11610.
43. Son, J.H., Xie, G., Yuan, C., Ena, L., Li, Z., Goldstein, A., Huang, L., Wang, L., Shen, F., Liu, H., et al. (2018). Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes. *Am J Hum Genet* 103, 58-73. 10.1016/j.ajhg.2018.05.010.
44. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., and Gray, A. Training language models to follow instructions with human feedback.
45. Fischer-Zirnsak, B., Segebrecht, L., Schubach, M., Charles, P., Alderman, E., Brown, K., Cadieux-Dion, M., Cartwright, T., Chen, Y., Costin, C., et al. (2019). Haploinsufficiency of the Notch Ligand DLL1 Causes Variable Neurodevelopmental Disorders. *Am J Hum Genet* 105, 631-639. 10.1016/j.ajhg.2019.07.002.
46. Maia, N., Potelle, S., Yildirim, H., Duvet, S., Akula, S.K., Schulz, C., Wiame, E., Gheldof, A., O'Kane, K., Lai, A., et al. (2022). Impaired catabolism of free oligosaccharides due to MAN2C1 variants causes a neurodevelopmental disorder. *Am J Hum Genet* 109, 345-360. 10.1016/j.ajhg.2021.12.010.
47. Yap, Z.Y., Efthymiou, S., Seiffert, S., Vargas Parra, K., Lee, S., Nasca, A., Maroofian, R., Schrauwen, I., Pendziwiat, M., Jung, S., et al. (2021). Bi-allelic variants in OGDHL cause a neurodevelopmental spectrum disease featuring epilepsy, hearing loss, visual impairment, and ataxia. *Am J Hum Genet* 108, 2368-2384. 10.1016/j.ajhg.2021.11.003.
48. Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., and Wallace, B. (2021). Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? held in Online, June. (Association for Computational Linguistics), pp. 946-959.
49. Hu, Y., Jing, X., Ko, Y., and Rayz, J.T. (2020). Misspelling Correction with Pre-trained Contextual Language Model. 2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), 144-149.
50. Chambon, P.J., Wu, C., Steinkamp, J.M., Adleberg, J., Cook, T.S., and Langlotz, C.P. (2023). Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. *J Am Med Inform Assoc* 30, 318-328. 10.1093/jamia/ocac219.
51. Clark, J., Garrette, D., Turc, I., and Wieting, J. (2021). Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics* 10, 73-91.
52. Dai, Z., Lai, G., Yang, Y., and Le, Q.V. (2020). Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing. *ArXiv abs/2006.03236*.

53. Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G., et al. (2022). A large language model for electronic health records. *NPJ Digit Med* 5, 194. 10.1038/s41746-022-00742-2.
54. Penedo, G., Malartic, Q., Hesslow, D., Cojocar, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116.
55. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., and Azhar, F. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
56. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Transformers: State-of-the-Art Natural Language Processing.
57. Organization, W.H. (1993). The ICD-10 classification of mental and behavioural disorders (World Health Organization).
58. Yan, S., Luo, L., Lai, P.-T., Veltri, D., Oler, A.J., Xirasagar, S., Ghosh, R., Similuk, M., Robinson, P.N., and Lu, Z. (2022). PhenoRerank: A re-ranking model for phenotypic concept recognition pre-trained on human phenotype ontology. *Journal of biomedical informatics* 129, 104059.
59. Groza, T., Köhler, S., Doelken, S., Collier, N., Oellrich, A., Smedley, D., Couto, F.M., Baynam, G., Zankl, A., and Robinson, P.N. (2015). Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database* 2015.
60. Anazi, S., Maddirevula, S., Salpietro, V., Asi, Y.T., Alsahli, S., Alhashem, A., Shamseldin, H.E., AlZahrani, F., Patel, N., and Ibrahim, N. (2017). Expanding the genetic heterogeneity of intellectual disability. *Human genetics* 136, 1419-1429.
61. Kudo, T., and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. pp. 66-71.
62. Taboada, M., Rodriguez, H., Martinez, D., Pardo, M., and Sobrido, M.J. (2015). Automated semantic annotation of rare disease cases: a case study. *Database (Oxford)* 2015. 10.1093/database/bav107.
63. Lison, P., Barnes, J., Hubin, A., and Touileb, S. (2020). Named Entity Recognition without Labelled Data: A Weak Supervision Approach. pp. 1518-1533.
64. Jiang, H., Zhang, D., Cao, T., Yin, B., and Zhao, T. (2021). Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data. pp. 1775-1789.

65. Yoon, W., Yi, S.S., Jackson, R., Kim, H., Kim, S., and Kang, J. (2021). Using Knowledge Base to Refine Data Augmentation for Biomedical Relation Extraction KU-AZ team at the BioCreative 7 DrugProt challenge.
66. Jiang, H. (2021). Reducing Human Labor Cost in Deep Learning for Natural Language Processing. (Georgia Institute of Technology).
67. Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., and Elhage, N. (2022). Predictability and surprise in large generative models. pp. 1747-1764.
68. Gupta, P., Pagliardini, M., and Jaggi, M. (2019). Better Word Embeddings by Disentangling Contextual n-Gram Information. In CONF. pp. 933–939.
69. Honnibal, M., and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
70. Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. (2022). Llm.int8(): 8-bit matrix multiplication for transformers at scale. arXiv preprint arXiv:2208.07339.
71. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
72. Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.

Legends of Figures

Figure 1. Illustration of the workflow of the project.

Figure 2. Illustration of the BERT-based and GPT-based model used in the current study for a sentence with phenotype mention. (A) Conversion of input sequence into a combination of three embeddings (sentence embedding is treated as zero in the current study). (B) The pre-training and fine-tuning strategy for PhenoBCBERT. (C) The pre-training and fine-tuning strategy for PhenoGPT.

Figure 3. Examples of phenotype terms from 8 clinical notes recognized by PhenoTagger and PhenoBCBERT. The font size is relative to the frequency of appearance.

Figure 4. Examples of phenotype terms from clinical notes recognized by different series of GPT models. GPT comparison 1: prediction results after prompt-based learning. GPT comparison 2-6: prediction results after fine-tuning. GPT-3 (N): GPT-3 fine-tuned based on N instances.

Figure 5. Case studies of the predicted phenotype entities with PhenoTagger, PhenoBCBERT and PhenoGPT (GPT-3). The negation terms and misspelled terms from the original published manuscript are highlighted.

TABLES

	Rule-based	Hybrid	Deep-learning	Year
Metamap	✓			2001
NCBO	✓			2009
OBO	✓			2014
Doc2Hpo	✓			2019
ClinPhen	✓			2019
NCR		✓	CNN (local)	2019
Phenotagger		✓	BERT(local)	2021
BERN2			BERT	2022
PhenoBERT		✓	CNN+BERT	2022

Table 1. Summary of different phenotype recognition models.

Input	BERT Tokenize	GPT Tokenize	Labeling
	[CLS]		-100
he	he	he	0
has	has	_has	0
red	red	_red	1
eyes	eyes	_eyes	1
and	and	_and	0
distance	distance	_distance	1
exotropia	exo	_exo	1
	##tropia	tropia	1
	[SEP]		-100

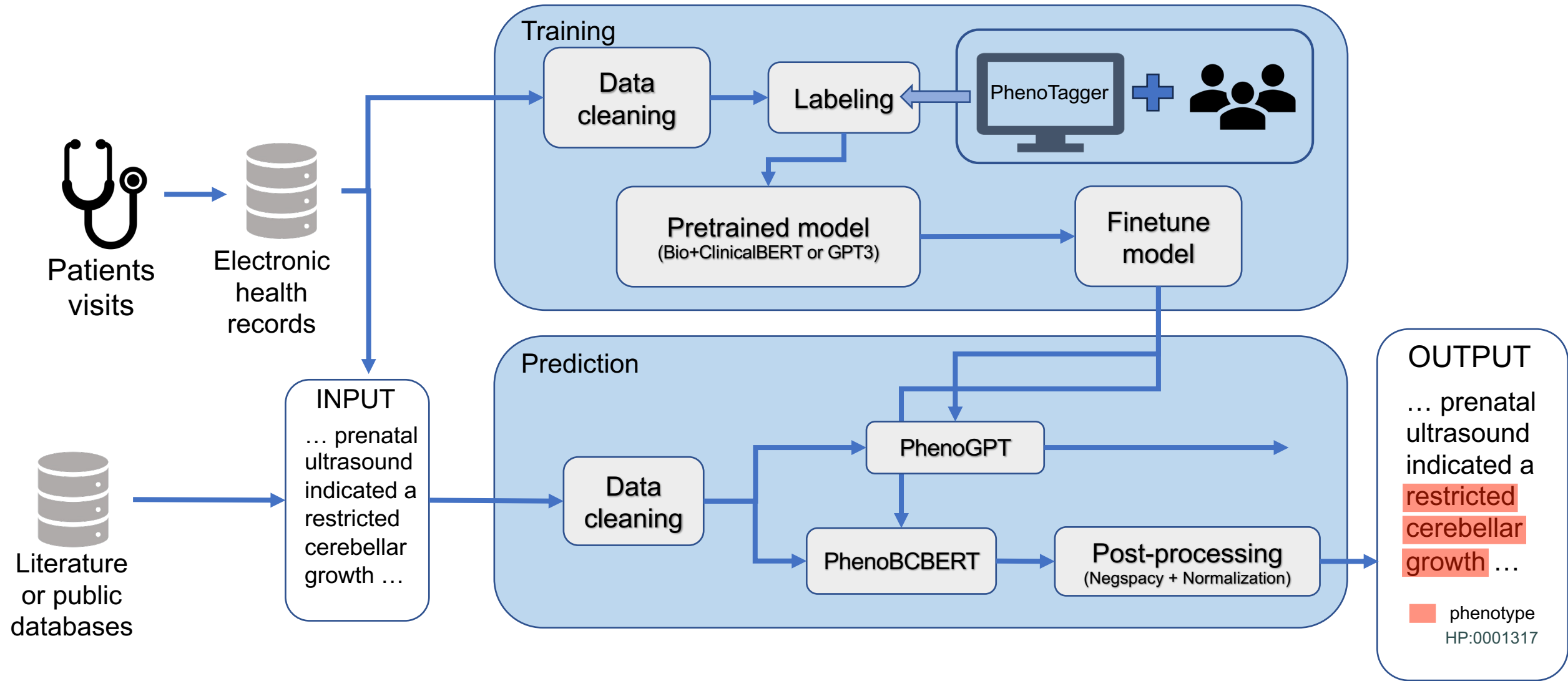
Table 2. Illustration of labeling strategies for a sentence with phenotype mentions.

	BiolarkGSC+ (Validation)		
Model	Precision	Recall	F1
OBO Anotator	0.810	0.568	0.668
NCBO	0.777	0.521	0.624
Doc2hpo-Ensemble	0.754	0.608	0.673
MetaMap	0.707	0.599	0.649
Clinphen	0.590	0.418	0.489
NeuralCR	0.736	0.610	0.667
PhenoTagger	0.720	0.760	0.740
PhenoRerank	0.843	0.708	0.770
PhenoBCBERT	0.747	0.813	0.779
PhenoGPT(GPT-3)	0.827	0.794	0.810
PhenoGPT(GPT-J)	0.809	0.857	0.832
PhenoGPT(Falcon)	0.801	0.792	0.796
PhenoGPT(LLaMA)	0.828	0.694	0.755

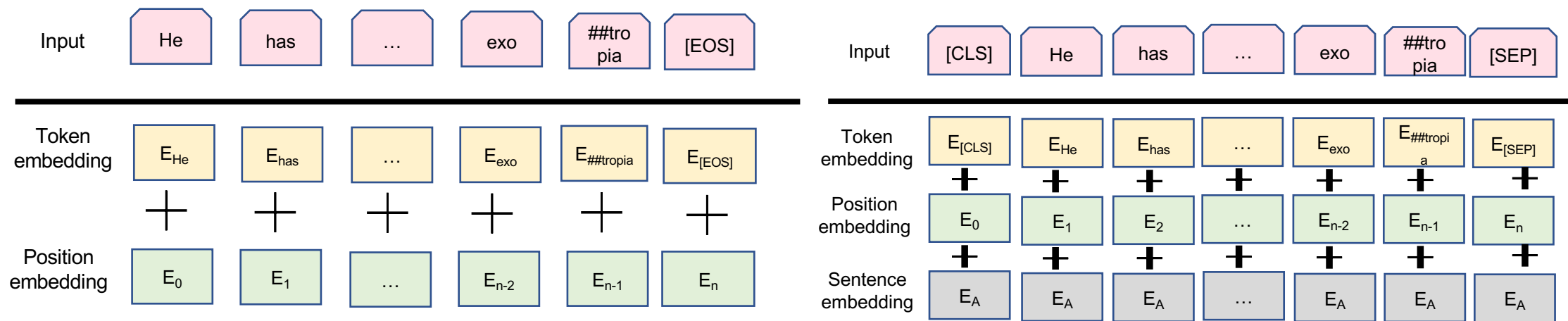
Table 3. Performance comparison on the BiolarkGSC+ validation set. PhenoGPT and PhenoBCBERT were fine-tuned using the BiolarkGSC+ training set.

	ID-68 (Validation)		
Model	Precision	Recall	F1
NCBO	0.874	0.660	0.752
MetaMapLite	0.804	0.591	0.682
Doc2hpo	0.844	0.575	0.684
Clinphen	0.749	0.615	0.675
NeuralCR	0.786	0.776	0.781
PhenoTagger	0.898	0.755	0.820
PhenoBERT	0.943	0.781	0.854
PhenoBCBERT	0.912	0.923	0.872
PhenoGPT(GPT-3)	0.818	0.814	0.816
PhenoGPT(GPT-J)	0.723	0.758	0.740
PhenoGPT(Falcon)	0.738	0.881	0.803
PhenoGPT(LLaMA)	0.719	0.926	0.809

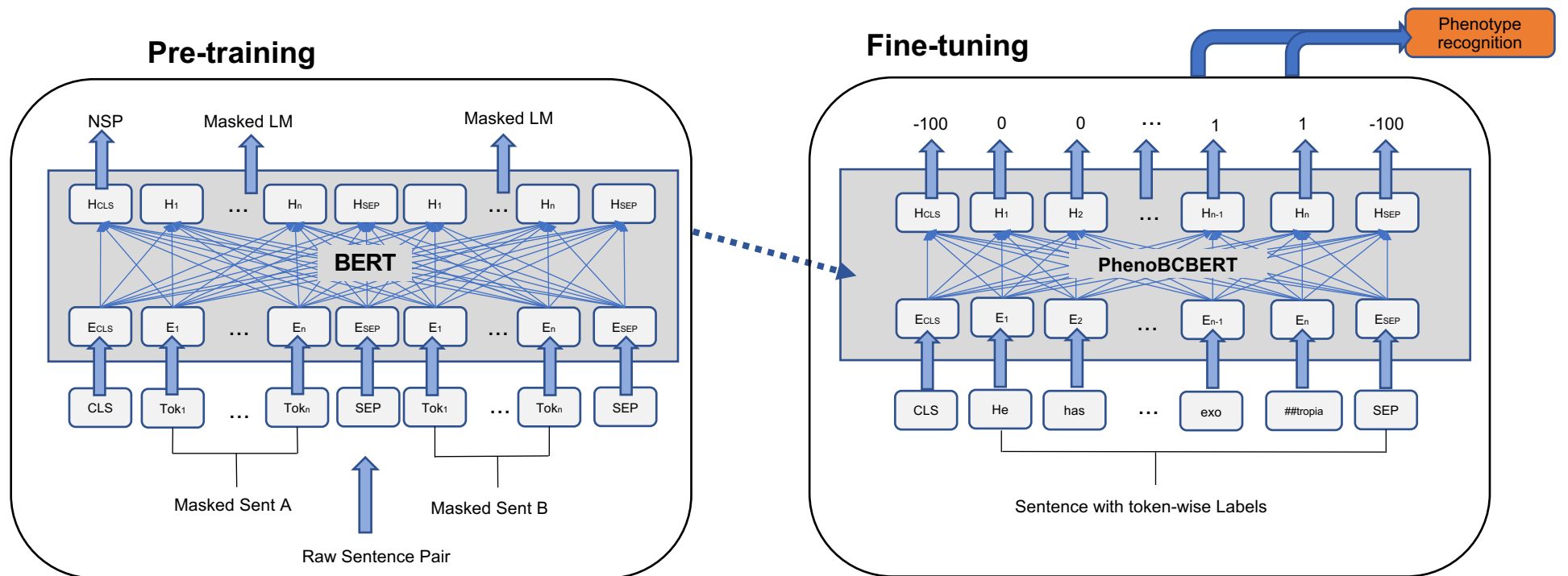
Table 4. Performance comparison on the ID-68 dataset. PhenoGPT and PhenoBCBERT were fine-tuned using BiolarkGSC+ training set.



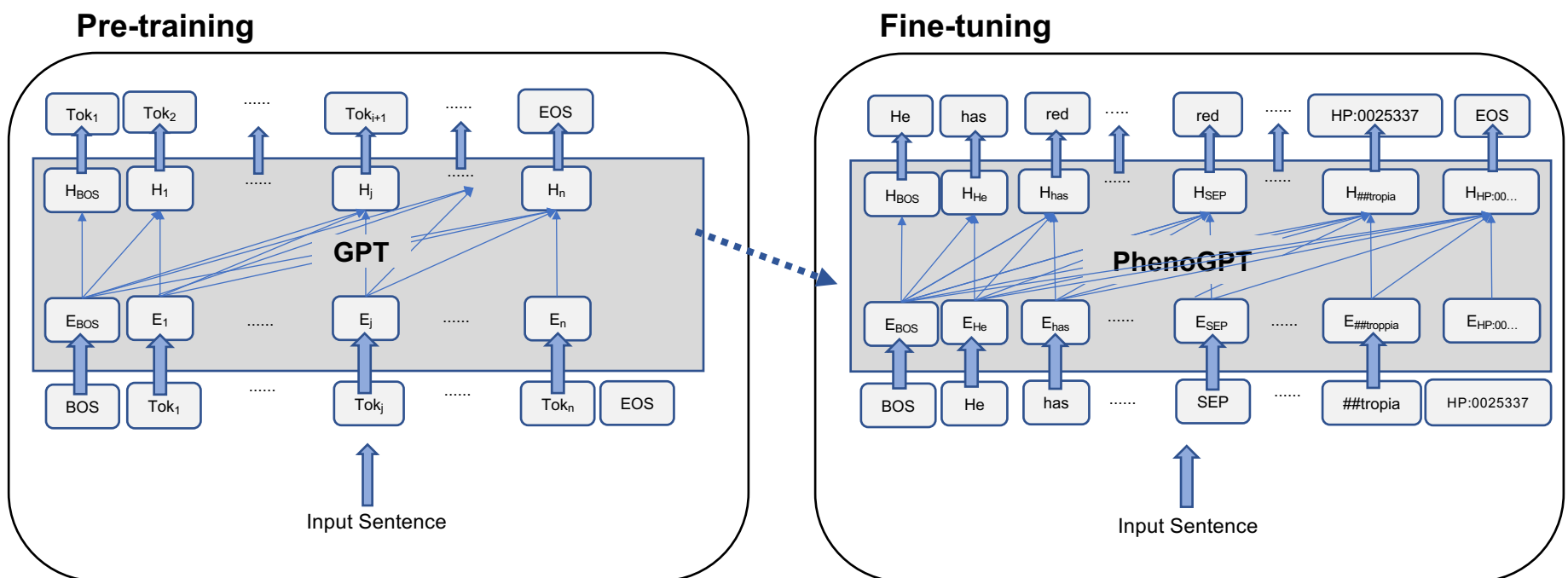
A

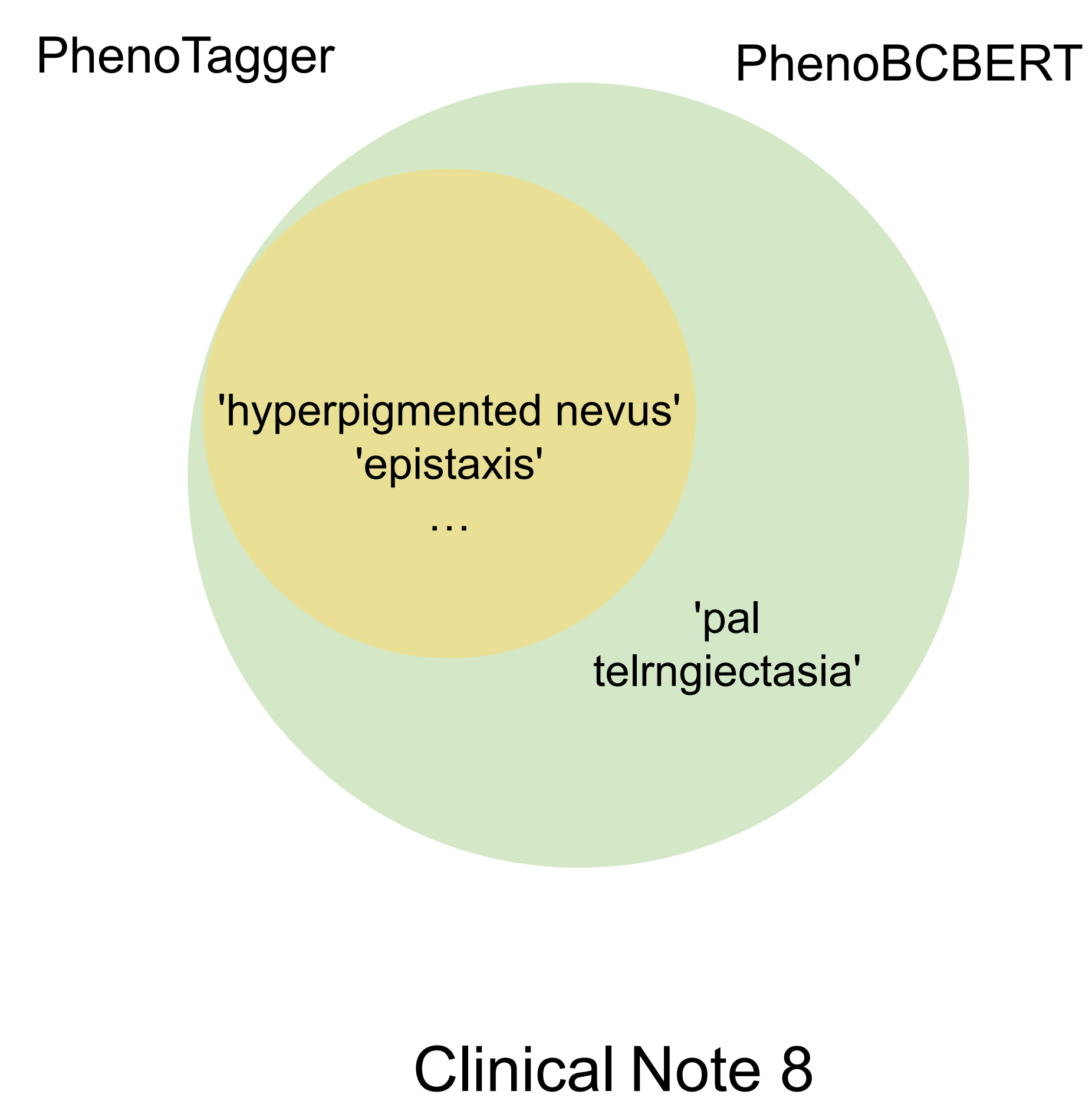
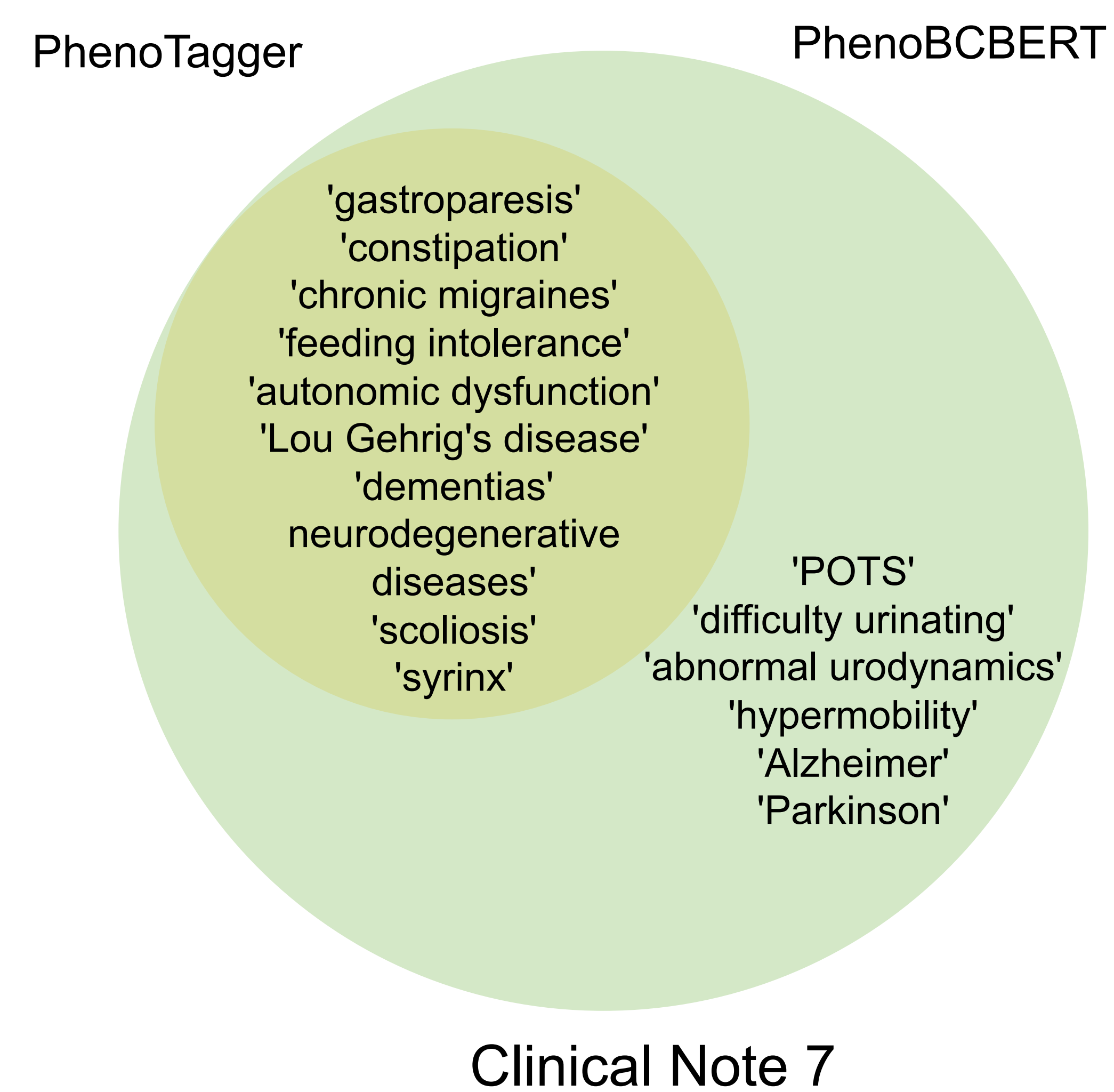
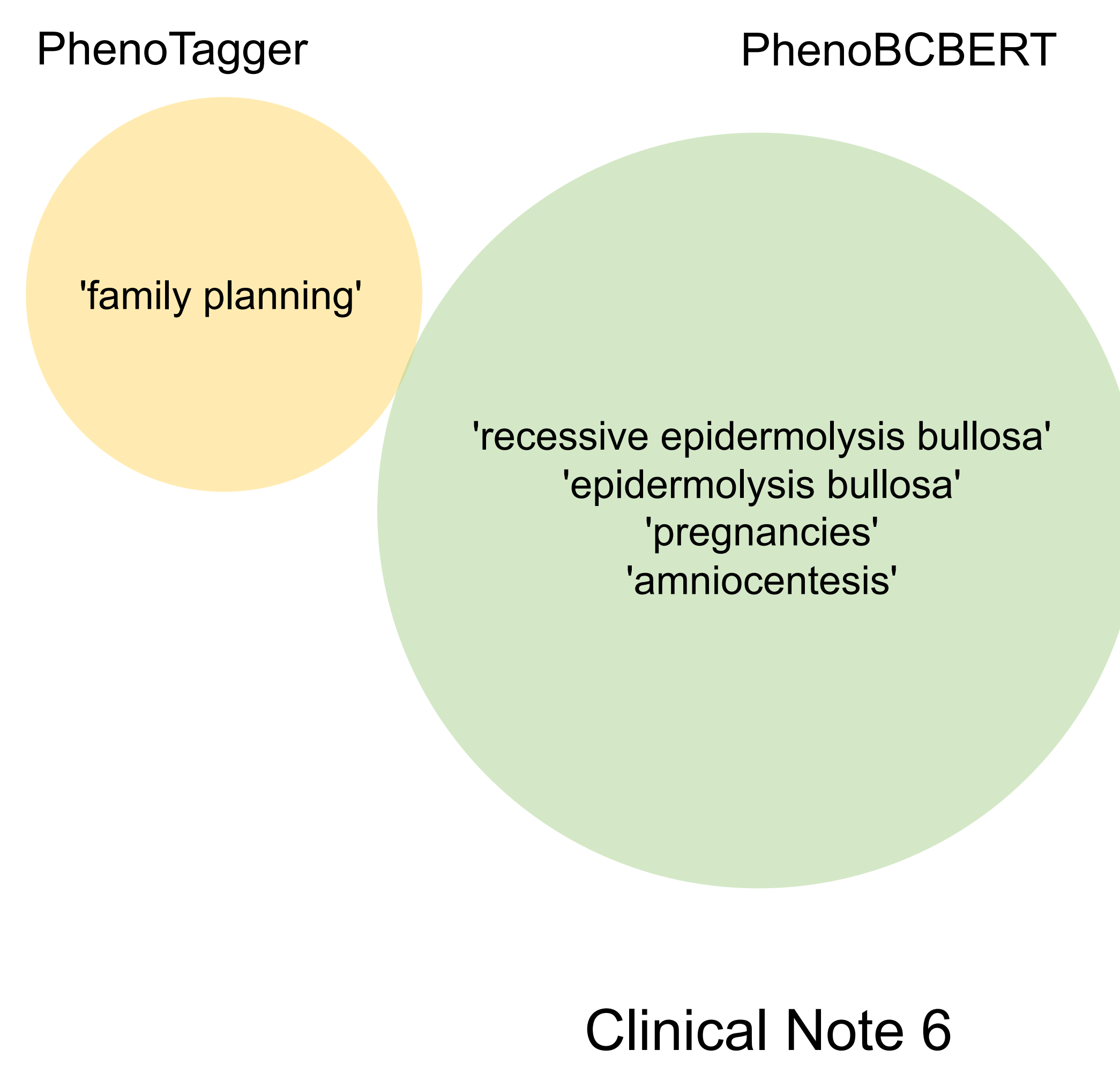
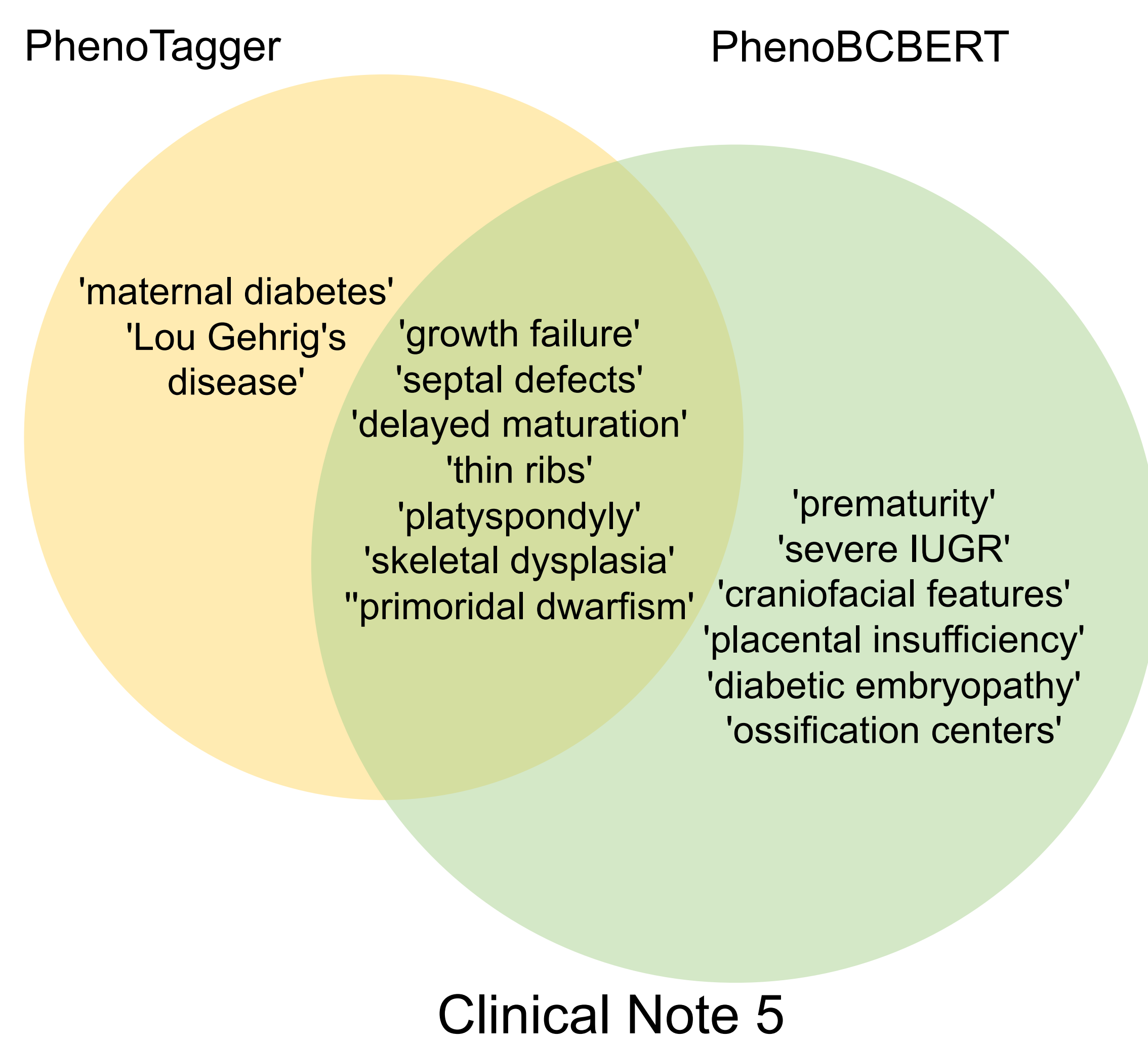
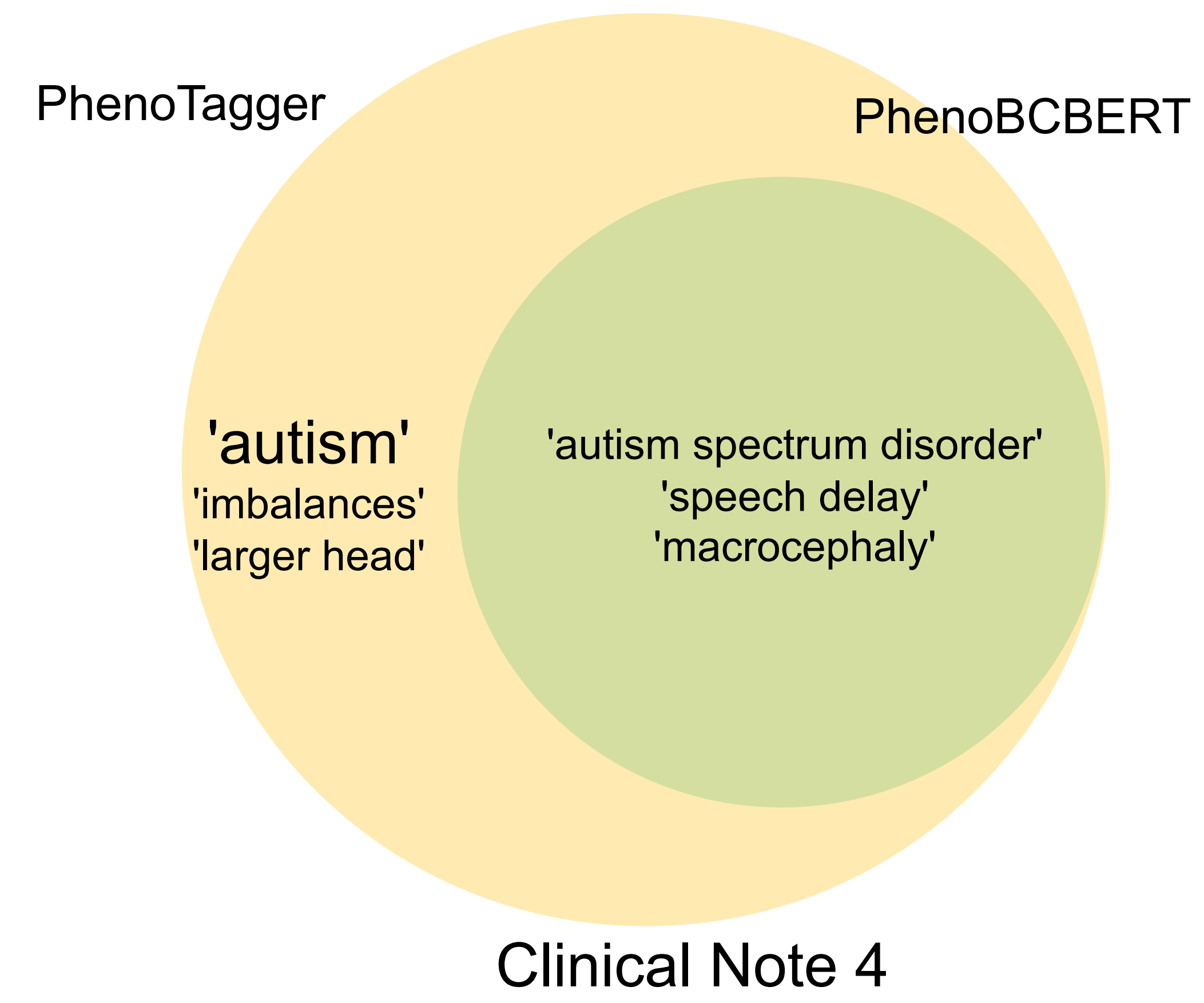
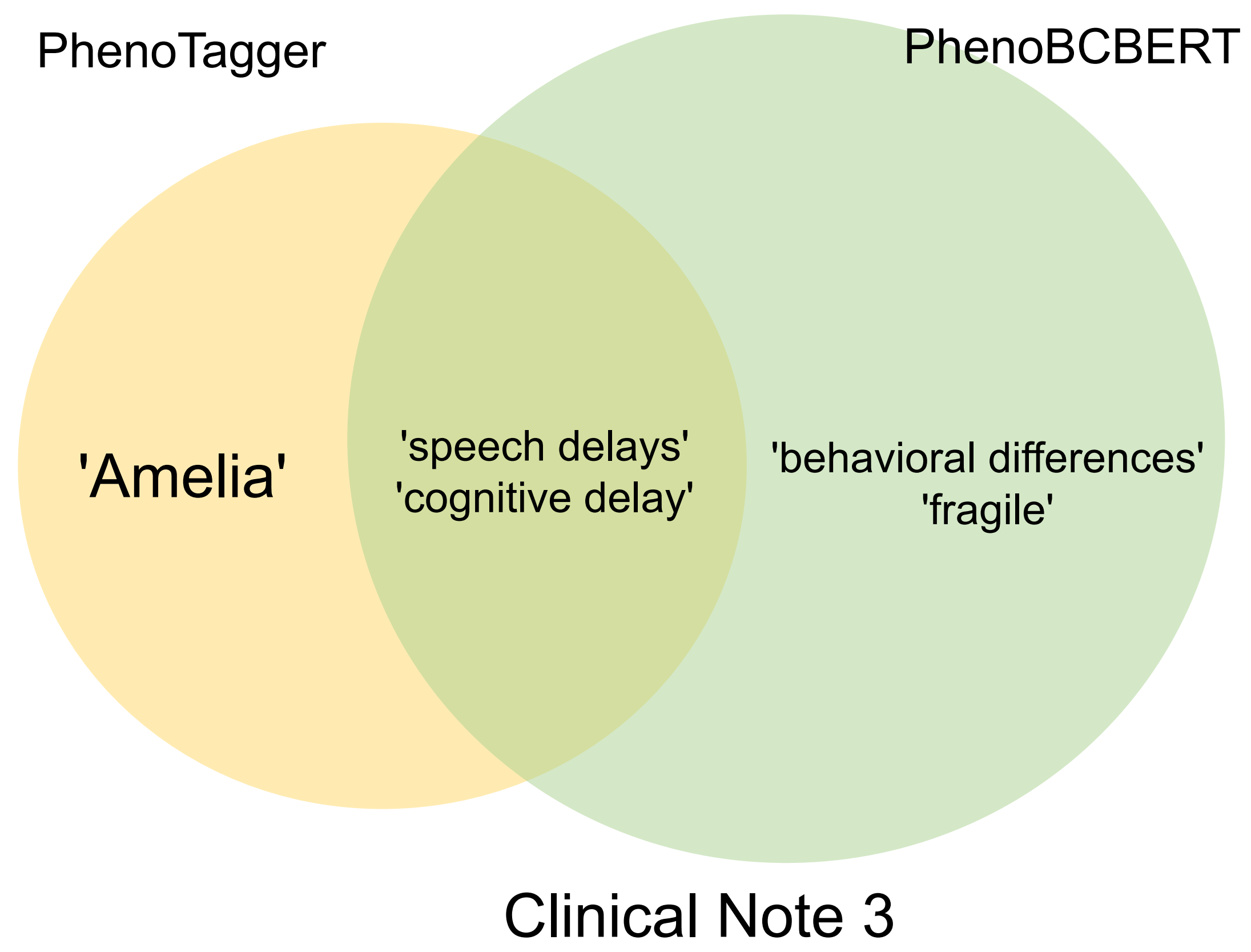
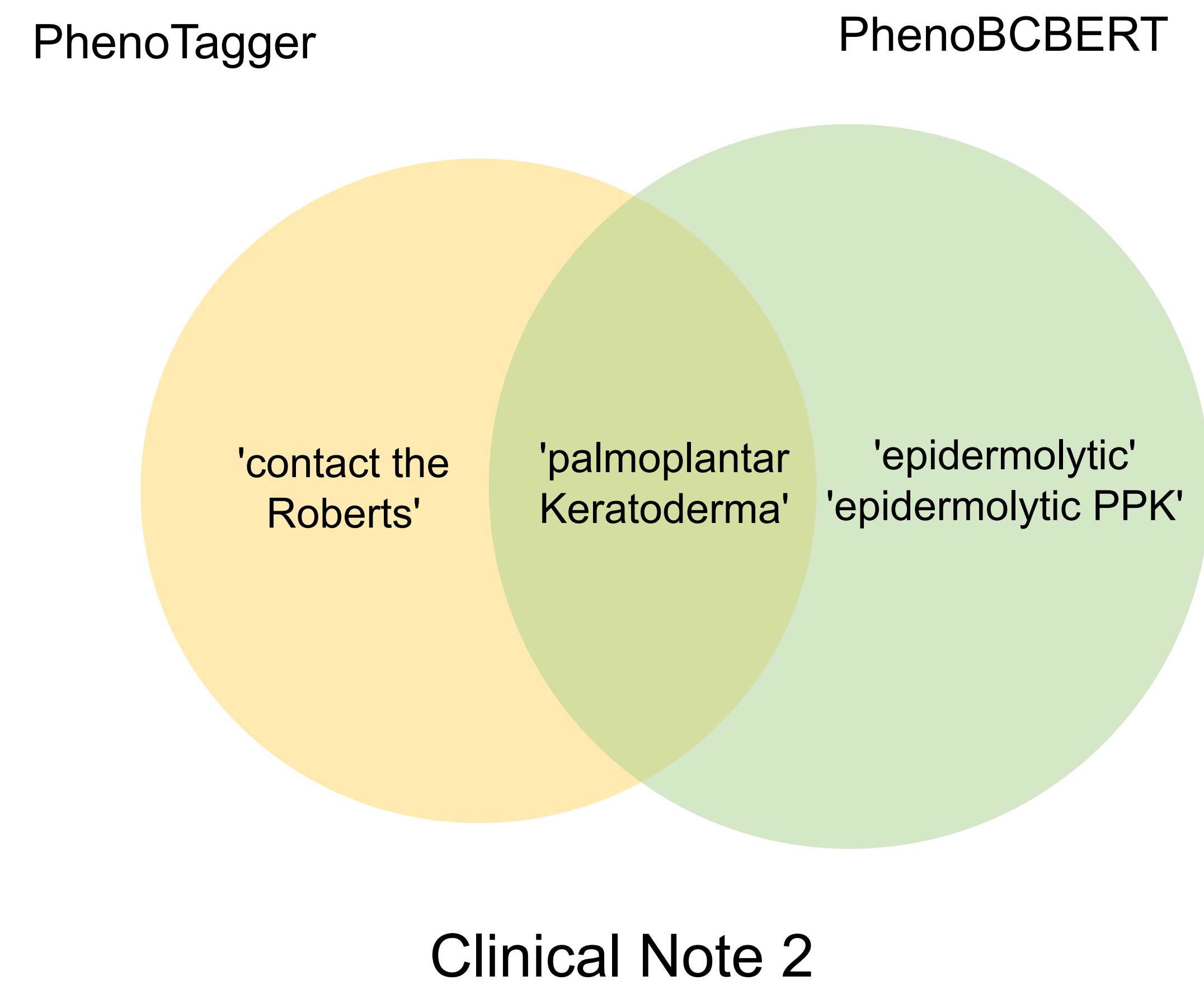
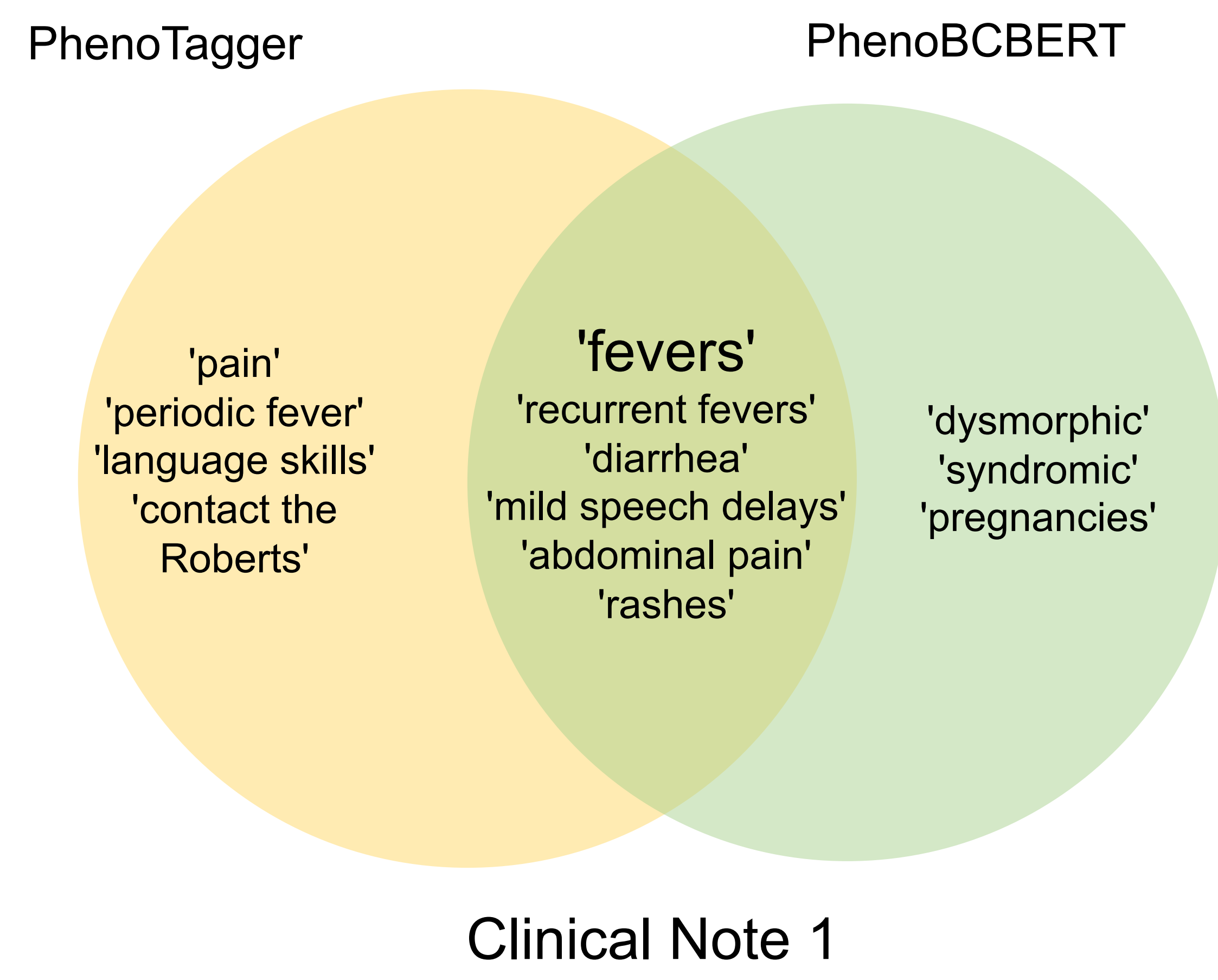


B

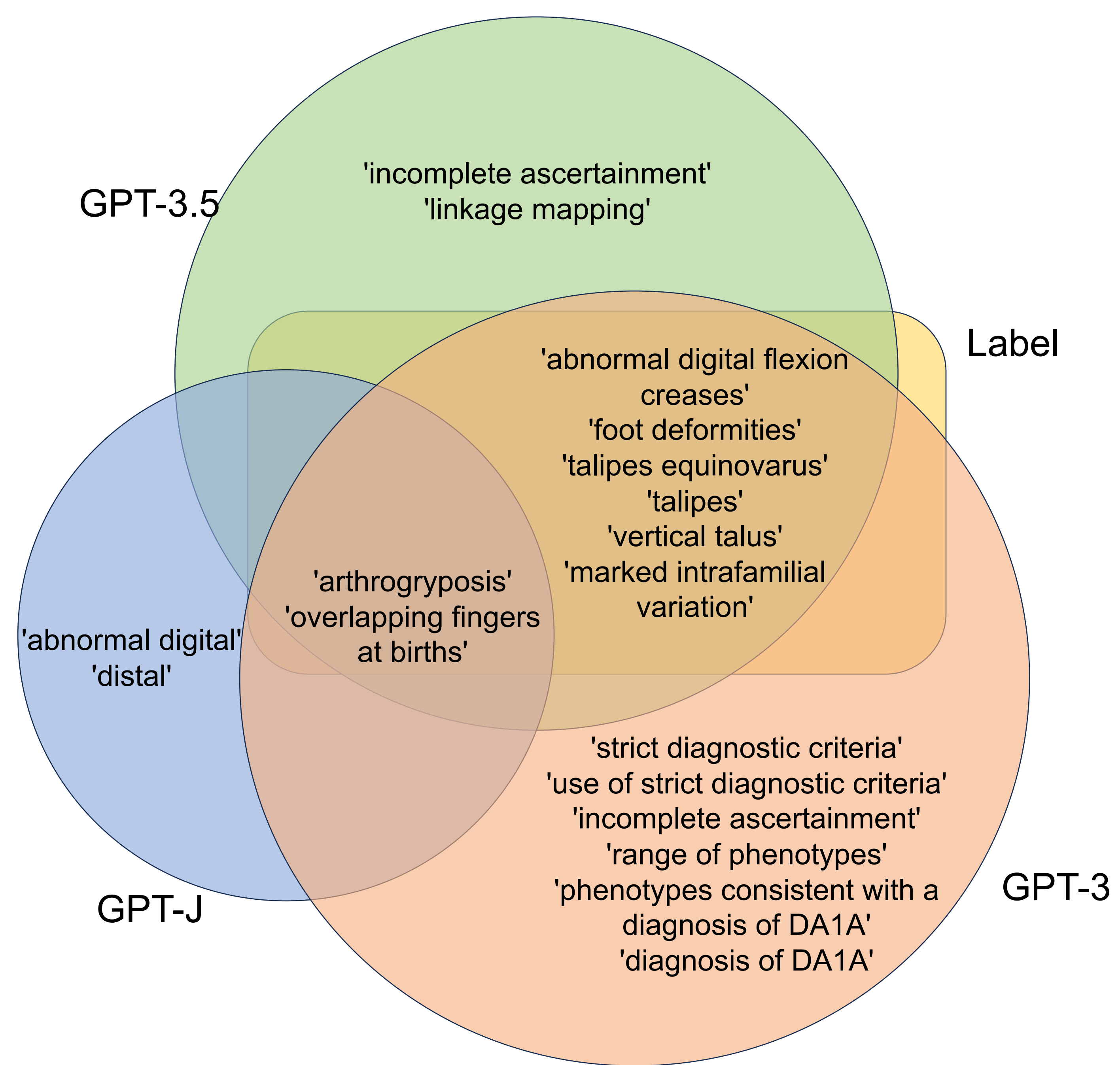


C

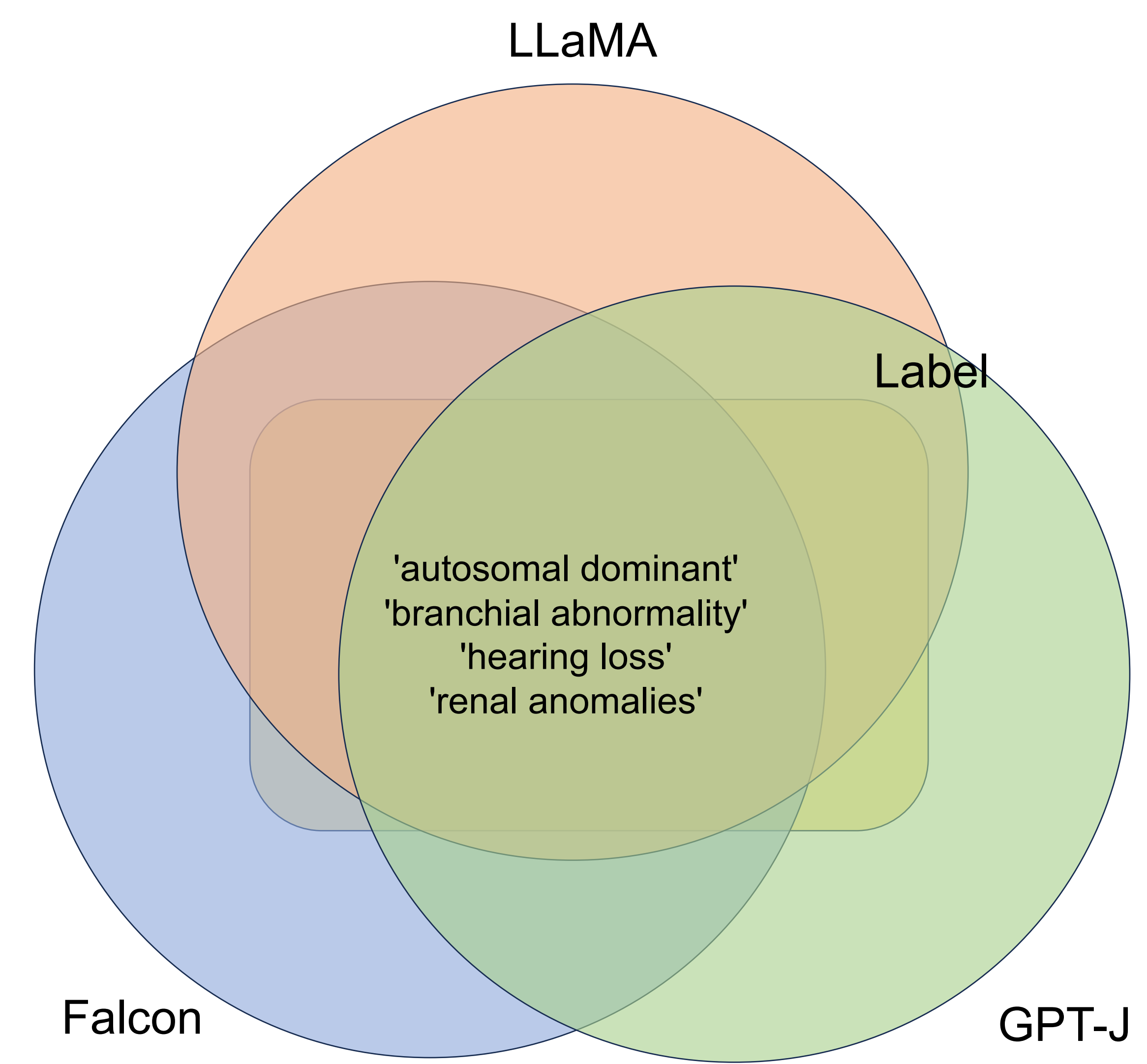




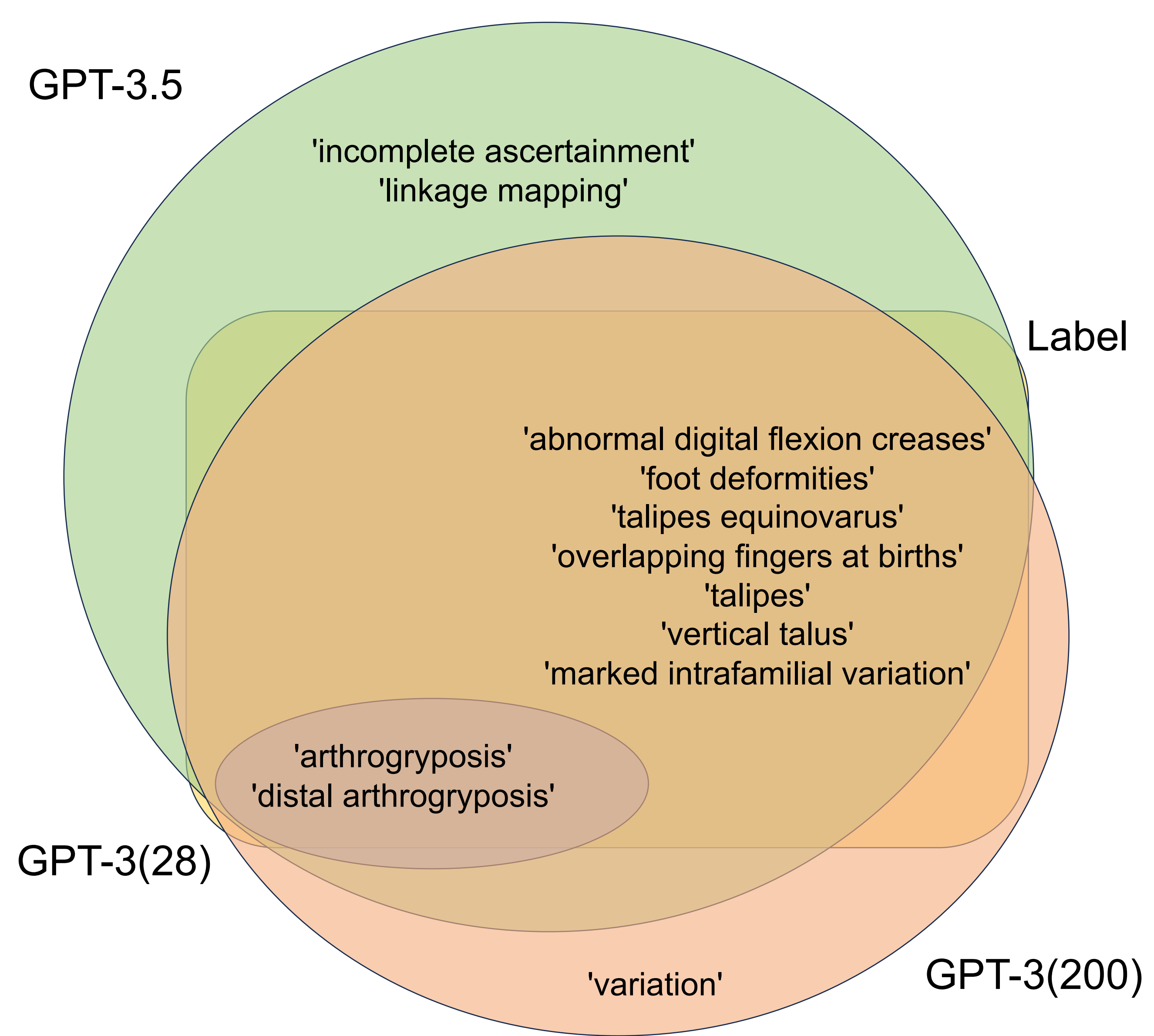
GPT comparison 1



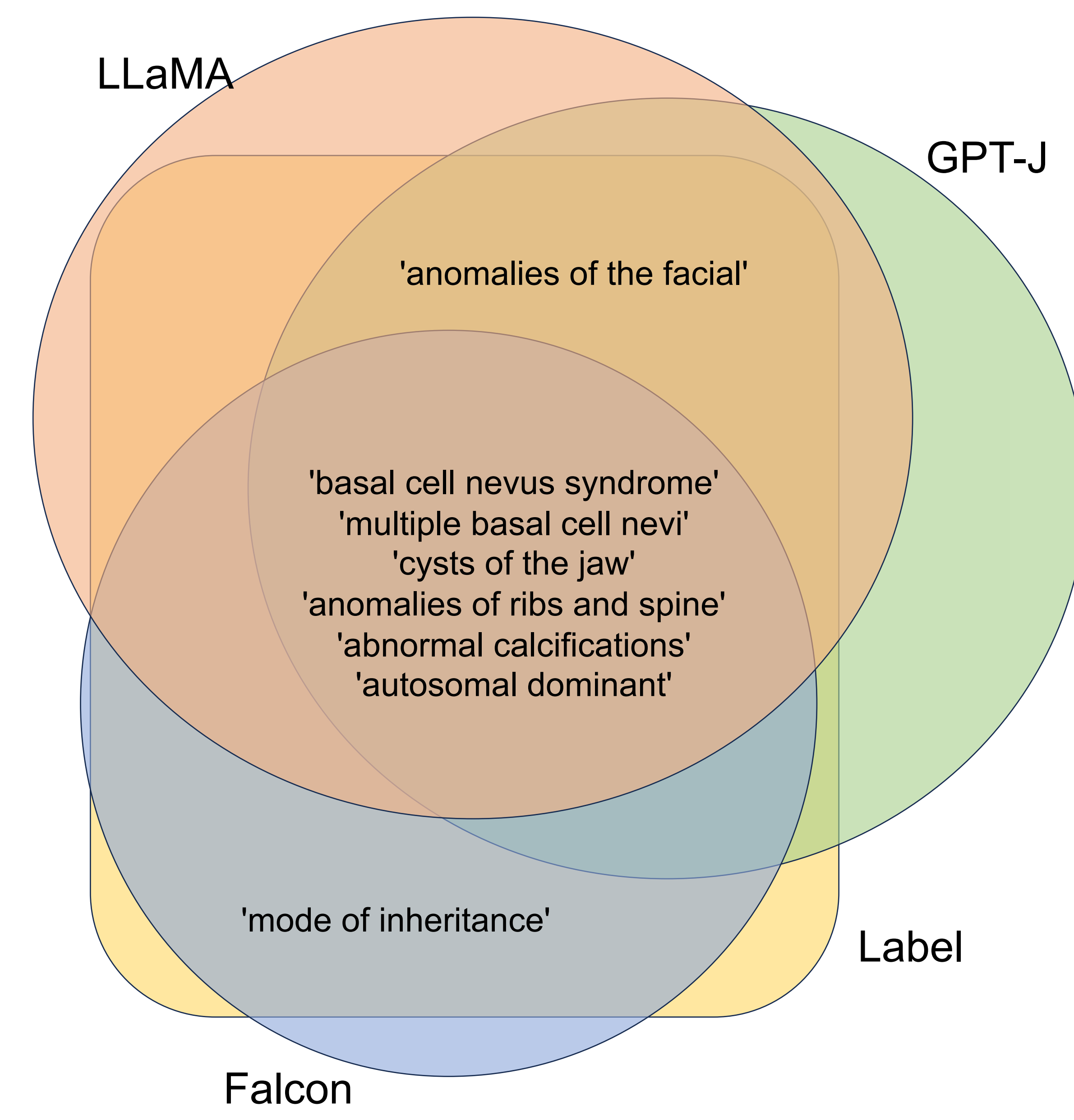
GPT comparison 4



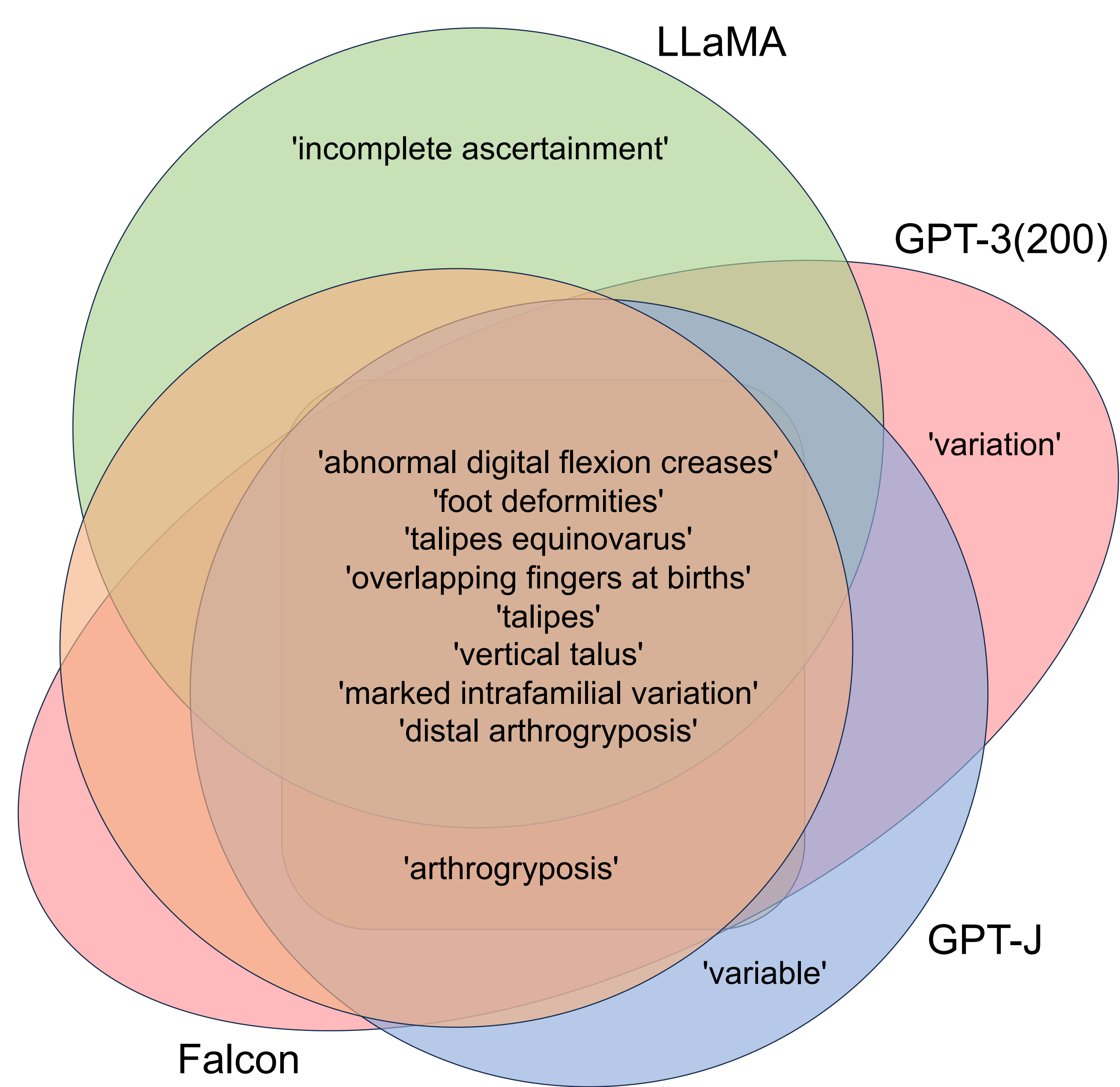
GPT comparison 2



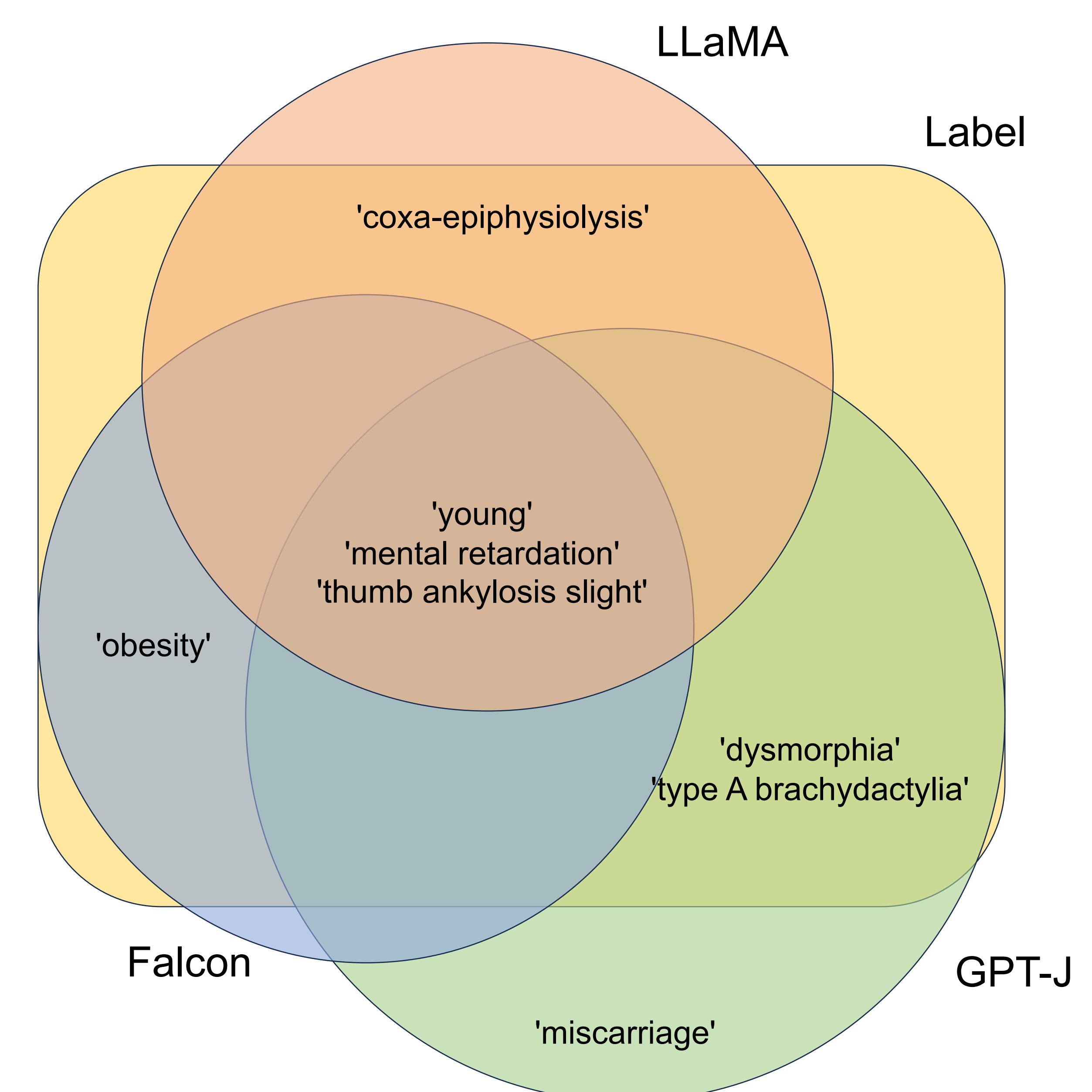
GPT comparison 5



GPT comparison 3



GPT comparison 6



- A. Individual 1 is a three-year-old female with developmental delays in expressive speech. Prenatal ultrasound indicated a restricted cerebellar growth and fetal magnetic resonance imaging (MRI) showed ventriculomegaly. A small cerebellum was also reported, but was not seen on review of the images. Prenatal triple screen, nuchal translucency and non-invasive prenatal testing (Harmony) did not indicate chromosomal abnormalities. Labor was noticeable for prolonged rupture of membranes. After birth, she had hyperbilirubinemia, which resolved with sunlight. She had mild muscular hypotonia. She was breastfed, but had problems with latching and swallowing.
- B. Two siblings were referred to the genetics clinic for mild global developmental delay. Individual 1 was born at term following vacuum extraction. Biometry and birth parameters were normal. Pregnancy was complicated as a result of gestational diabetes and the finding of a single umbilical artery on fetal ultrasound. Head circumference was normal, but height and weight were above the 95th centile as of age 18 and 24 months and beyond, respectively. Renal ultrasound revealed a pyelocaliceal dilatation, which regressed spontaneously. Evaluation at age 7 years showed normal intellectual ability, specific learning difficulties mainly related to expressive language, attention deficit hyperactivity disorder, poor fine motor skills, and mild non-specific dysmorphisms including high forehead, arched eyebrows, short columella, micrognathia, and small ears.
- C. The first symptoms started at 8 years of age when a severe neurosensory hyopoacusia was found. At 9 years, he developed gait difficulty with stiffness. At the same age, he started to present a dorsal right-convex scoliosis and a retinopathy was detected. Neurological examinations at 16 years showed muscle weakness and diffuse muscle hypotrophy, more severe in limbs and distally. Pyramidal signs such as high tendon reflexes (Babinski and Achilles tendon clonus) were detected and muscle tone was high in limbs. Pes cavus was observed bilaterally, although more evident in the right foot. Achilles tendon retraction was present. The individual presented with distal tremor.

PhenoTagger —
 PhenoBCBERT —
 PhenoGPT —
 Negation —
 Misspelling —