frontiers in Digital Health

TYPE Original Research
PUBLISHED 21 May 2026
DOI 10.3389/fdgth.2026.1794934

OPEN ACCESS

EDITED BY
Mohan Bhandari,
Samridhhi College, Nepal

REVIEWED BY
Mayada Abu Shanap,
King Hussein Cancer Center, Jordan
Eisuke Dohi,
National Center of Neurology and Psychiatry, Japan

*CORRESPONDENCE
Kitty B. Murphy
c.murphy19@imperial.ac.uk
Nathan G. Skene
n.skene@imperial.ac.uk

RECEIVED 24 January 2026
REVISED 10 April 2026
ACCEPTED 13 April 2026
PUBLISHED 21 May 2026

CITATION
Murphy KB, Schilder BM and Skene NG (2026) Using GPT-4 to annotate the severity of all phenotypic abnormalities within the human phenotype ontology. Front. Digit. Health 8:1794934. doi: 10.3389/fdgth.2026.1794934

Using GPT-4 to annotate the severity of all phenotypic abnormalities within the human phenotype ontology

Kitty B. Murphy[1,2]*, Brian M. Schilder[1,2,3,4] and Nathan G. Skene[1,2]*

[1]Department of Brain Sciences, Imperial College London, London, United Kingdom, [2]UK Dementia Research Institute, Imperial College London, London, United Kingdom, [3]Icahn School of Medicine at Mount Sinai, New York, NY, United States, [4]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, United States

Introduction: The Human Phenotype Ontology (HPO) provides a unified framework cataloguing over 17,500 phenotypic abnormalities across more than 8,600 rare diseases, defining hierarchical relationships between them. For example, classifying missing arms and missing legs as both abnormalities of the limb. This structure enables phenome-wide analyses, including the prioritisation of phenotypes as candidates for gene therapy. However, the HPO currently lacks sufficient metadata describing the clinical severity of these phenotypes. Manual expert curation at this scale would be prohibitively labour-intensive, creating a need for automated approaches to systematically annotate phenotypic severity.
Methods: GPT-4, a large language model (LLM) developed by OpenAI, was employed to annotate the severity of all phenotypic abnormalities catalogued in the HPO. Severity was operationalised using nine clinical characteristics: congenital onset, reduced fertility, sensory impairments, impaired mobility, immunodeficiency, physical malformations, cancer, intellectual disability, and death. Each characteristic was further qualified by frequency of occurrence across four levels: never, rarely, often, and always. To assess annotation quality, GPT-4's outputs were benchmarked against ground-truth labels embedded within the HPO itself. For instance, phenotypes residing in the "Cancer" HPO branch were expected to be annotated as cancer-causing. A novel severity scoring system was then developed that integrates both the nature of each clinical characteristic and its frequency of occurrence.
Results: Benchmarking demonstrated strong performance across all clinical characteristics, with true positive recall rates ranging from 89% to 100% (mean = 97%). This indicates that GPT-4 can replicate expert-level curation with high fidelity. The resulting severity scoring system produced quantitative severity metrics for phenotypic abnormalities across the HPO, incorporating both the type and frequency of associated clinical characteristics.
Discussion: These findings demonstrate that LLMs can automate the large-scale curation of clinical metadata with a high degree of accuracy, substantially reducing the burden of manual expert annotation. The severity metrics generated here provide a foundation for systematically ranking human phenotypes by their impact on health and quality of life, enabling more principled prioritisation of targets for therapeutic intervention, particularly in the context of rare diseases where evidence is sparse and resources for curation are limited. Future work may extend this framework to incorporate additional clinical dimensions or validate annotations against independent clinical datasets.

KEYWORDS
artificial intelligence, generative AI, GPT-4, human phenotype ontology (HPO), large language mode1, medicine, rare disease

Introduction

Ontologies provide a common language with which to communicate concepts. In medicine, ontologies for phenotypic abnormalities are invaluable for defining, diagnosing, prognosing, and treating human disease. Since 2008, the Human Phenotype Ontology (HPO) has been instrumental in healthcare and biomedical research by providing a framework for comprehensively describing human phenotypes and the relationships between them (1, 2). By expanding its depth and breadth over time, the HPO now contains >17,500 phenotypic abnormalities that are associated with >8,600 diseases as captured in HPOA. Some HPO phenotypes also contain information related to typical age of onset, frequency, triggers, time course, mortality rate and typical severity. Describing the severity-related attributes of a disease is crucial for both research and clinical care of individuals with rare diseases. When researchers or clinicians are presented with phenotypes that fall outside of their expertise, resources to quickly and reliably retrieve summaries with additional relevant information about these phenotypes are essential. In the clinic, this can help in reaching a differential diagnosis or prioritising the treatment of some phenotypes over others. In research, this information is useful for prioritising targets for causal disease mechanisms, performing large-scale analyses of phenotypic data, and guiding funding agencies when assessing the potential impact and need for research in a given disease area. To date, the HPO has largely relied on manual curation by domain experts. While this approach can improve annotation quality and accuracy, it is both time-consuming and labour-intensive. As a result, less than 1% of terms within the HPO contain metadata such as time course and severity.

Artificial intelligence (AI) capabilities have advanced considerably in recent years, presenting new opportunities to integrate natural language processing technologies into assisting in the curation process. Specifically, there have recently been considerable advances in large language model (LLM) and their application to biomedical problems, in some cases performing as well or better than human clinicians on standardised medical exams and patient diagnosis tasks (3–14). Recent work has demonstrated that the Generative Pre-trained Transformer 4 (GPT-4) foundation model (15), when combined with strategic prompt engineering, can outperform even specialist LLMs that are explicitly fine-tuned for biomedical tasks (16). In a landmark achievement, GPT-4 was the first LLM to surpass a score of 90% in the United States Medical Licensing Examination (USML) (16).

Here, we have used GPT-4 to systematically annotate the severity of 17,502/17,548 (99.7%) phenotypic abnormalities within the HPO. Our severity annotation framework was adapted from previously defined criteria developed through consultation with clinicians (17). The authors consulted 192 healthcare professionals for their opinions on the relative severity of various clinical characteristics: they used this to create a system for categorising the severity of diseases. Briefly, each healthcare professional was sent a survey asking them to first rate how important a disease characteristic was for determining disease severity, and then to rate the severity of a set of given disease. Using the responses, the authors were able to categorise clinical characteristics into 4 "severity tiers". While characteristics such as shortened lifespan in infancy and intellectual disability were identified as highly severe and placed into tier 1, sensory impairment and reduced lifespan were categorised as less severe and placed into tier 4. Standardised metrics of severity allow clinicians to quickly assess the urgency of treating a given phenotype, as well as prognosing what outcomes might be expected. Here, we used the clinical characteristics introduced in Lazarin et al. (17): congenital onset, reduced fertility, sensory impairments, impaired mobility, immunodeficiency, physical malformations, cancer, intellectual disability, and death. In Lazarin et al. (17), these were ranked by healthcare professionals based on how important they were to determine disease severity. We expand on this work to annotate each HPO phenotype according to how frequently it causes each characteristic—never, rarely, often, or always. Guided by the four-tier severity system defined in Lazarin et al. (17), we then developed a quantifiable severity score for each phenotype. While we recognise that severity can be subjective and highly personal, we believe that our study provides a strong framework for how LLMs can be used to scale phenotypic annotation and prioritisation for clinical trials.

To evaluate the consistency of responses generated by GPT-4 793 phenotypes were annotated multiple times. For a subset of phenotypes with known expected clinical characteristics, true positive rates were calculated to assess recall. Additionally, based on the clinical characteristics and their occurrence, we have quantified the severity of each phenotype, providing an example of how these clinical characteristic annotations can be used to guide prioritisation of gene therapy trials. Ultimately, we hope that our resource will be of utility to those working in rare diseases, as well as the wider healthcare community.

Results

Annotating the HPO using GPT-4

We employed the OpenAI GPT-4 model with Python to annotate 17,502 terms within the HPO (v2024-02-08) (1, 2). Our annotation framework was developed based on previously defined criteria for classifying disease severity (17). We sought to evaluate the impact of phenotypes on factors including intellectual disability, death, impaired mobility, physical malformations, blindness, sensory impairments, immunodeficiency, cancer, reduced fertility, and congenital onset. Through prompt design we found that the performance of GPT-4 improved when we incorporated a scale associated with each clinical characteristic and required a justification for each response. For each clinical characteristic, we asked about the frequency of its occurrence—whether it never, rarely, often, or always occurred. Framing the queries in this way served two purposes. First, this helped to constrain the responses of GPT-4 to a specific range of values, making answers more consistent and amenable to downstream data analysis. Second, it served to overcome one of the main limitations noted by Lazarin et al. (17) as they did not collect information on how the frequency of each disease affected their decision making when generating severity annotations.

Clinical characteristic occurrence varied across annotation categories. >50% of phenotypes never caused blindness, sensory impairments, immunodeficiency, cancer, reduced fertility or intellectual disability. Only a minority of phenotypes (21.7%) never had a congenital onset, which is expected as rare disorders tend to be early onset genetic conditions (Figure 1).

Less than 1% of phenotypes always directly resulted in death ($n = 71$), such as "Stillbirth", "Anencephaly" and "Bilateral lung agenesis". Meanwhile, 45% of phenotypes were annotated as never causing death ($n = 7{,}880$). Examples of phenotypes that never cause death included 34 unique forms of syndactyly, a non-lethal condition that causes fused or webbed fingers (occurring 1 in 1,200–15,000 live births). Although syndactyly can occur alongside life-threatening conditions such as Aspert Syndrome (18), GPT-4 correctly annotated it as non-lethal, recgonising that the webbed fingers themselves do not cause death. This example highlights the model's ability to distinguish between phenotypes that directly cause lethality and those associated with diseases that do.

Annotation consistency and recall

To assess annotation consistency, we queried GPT-4 with a subset of the HPO phenotypes multiple times ($n = 793$ unique phenotypes). This subset was randomly selected and matches the ontology level distribution of all the phenotypes tested (Supplementary Figure S8). We employed two different metrics to determine the *consistency rate*. The first, less stringent metric, defined consistency as the duplicate annotations being either "always" and "often", or "never" and "rarely". The second, more stringent metric, required exact agreement in annotation occurrences, e.g., "always" and "always". For the less stringent metric, duplicated phenotypes were annotated consistently at a rate of at least 80%, and for the more stringent metric, the lowest consistency rate was 57% for congenital onset (Figure 2A). An example of where annotations were inconsistent was for the HPO term "Acute leukaemia". One time, GPT-4 annotated it as often causing impaired mobility, giving the justification that "weakness and fatigue from leukaemia and its treatment can impair mobility". The other time, GPT-4 annotated it as rarely causing impaired mobility, giving the justification that "acute leukaemia rarely impairs mobility directly". Despite specifying in the prompt for GPT-4 not to take into consideration indirect effects, this is an example of where it failed to do so. To investigate this further, clinical characteristic justifications were searched for terms indicating indirect causation ("indirectly", "does not directly"), identifying 4,495 HPO terms. The majority ($n = 3{,}516$) were classified as "never" directly causing the associated clinical characteristic, meaning that the indirect cause was not taken into account and only described in the justification. For example, "Obesity does not directly cause cancer, although it is a risk factor for various types of cancer". 1,385 terms rarely caused the associated clinical characteristic. For example, "While this condition (Third degree atrioventricular block) does not directly cause intellectual disability, complications such as cardiac arrest and insufficient oxygen delivery to the brain can cause damage leading to cognitive impairments". Only 20 phenotypes "often" caused the associated clinical characteristic, for example, "Increased C-peptide level is often associated with insulinoma, a rare pancreatic tumor, but does not directly cause it". 0 phenotypes "always" caused the associated clinical characteristic, when taking into account indirect causes. To assess the impact of this indirect causation on severity scoring, all indirect causal terms were reassigned to "never" and severity scores were recalculated. Although this resulted in a downward shift in raw severity scores, phenotype classifications into severity classes remained unchanged (Supplementary Figure S6).

We also reasoned that GPT-4 would be better able to give consistent answers for more specific phenotypes lower in the ontology, as they are more likely to have a single cause (Spearman's rank correlation rho, estimate = −0.254, $p = <2 \times 10^{-16}$), Supplementary Figure S7). We found that the stringent consistency rate did indeed significantly improve with greater HPO ontology depth ($X^2_{Pearson} = 22.17$, $\hat{V}_{Cramer} = 0.03$, $p = 0.05$). See Supplementary Figure S9 for a visual representation of this relationship.

In order to evaluate the validity of the annotations, we calculated a true positive rate. This involved identifying specific branches within the HPO that would contain phenotypes that would reliably indicate the presence of certain conditions. For instance, the phenotypes "Decreased fertility in females" and "Decreased fertility in males" should often or always cause reduced fertility. We observed an encouraging true positive rate exceeding 88% across in every clinical characteristic and achieving perfect recall (100%) in 5/9 characteristics (Figure 2B).

The lowest true positive rate was observed for physical malformations, with 88.5% recall across 87 HPO phenotypes. Some cases in which the GPT-4 annotations disagreed with the HPO ground truth included: "Angioma serpentinum", "Nevus anemicus", "Pulmonary arteriovenous fistulas". In the case of "Angioma serpentinum" it provided the justification that "No known association with physical malformations". In another instance, GPT-4 noted that "Nevus anemicus" is "Limited to hypopigmented skin patch; no other malformations.". This indicates that while technically incorrect according to our predefined benchmarks, a case could in fact be made that mild skin conditions do not rise to the level of physical malformations.

This high level of recall underscores the robustness of our annotations and the reliability of the HPO framework in capturing clinically relevant phenotypic information. However, we acknowledge that the number of testable true positive phenotypes for some of these categories are low, especially "blindness" for which there is only 1 phenotype in the HPO (after excluding terms pertaining to colour or night blindness, as these are not typically classified as causes of complete blindness). Furthermore, some of the true positive phenotypes are lexically similar to the name of the clinical characteristic itself. In these cases, annotating "Severe intellectual disability" as always causing intellectual disability is a relatively trivial task. Nevertheless, even these scenarios provide a clear and interpretable benchmark for the model's performance. In addition, there were numerous phenotypes with lexically non-obvious relationships to the clinical characteristic that were annotated correctly by GPT-4. For example, "Molar tooth sign on MRI" (a neurodevelopmental pathology observed in radiological scans) was correctly annotated as causing

10.3389/fdgth.2026.1794934

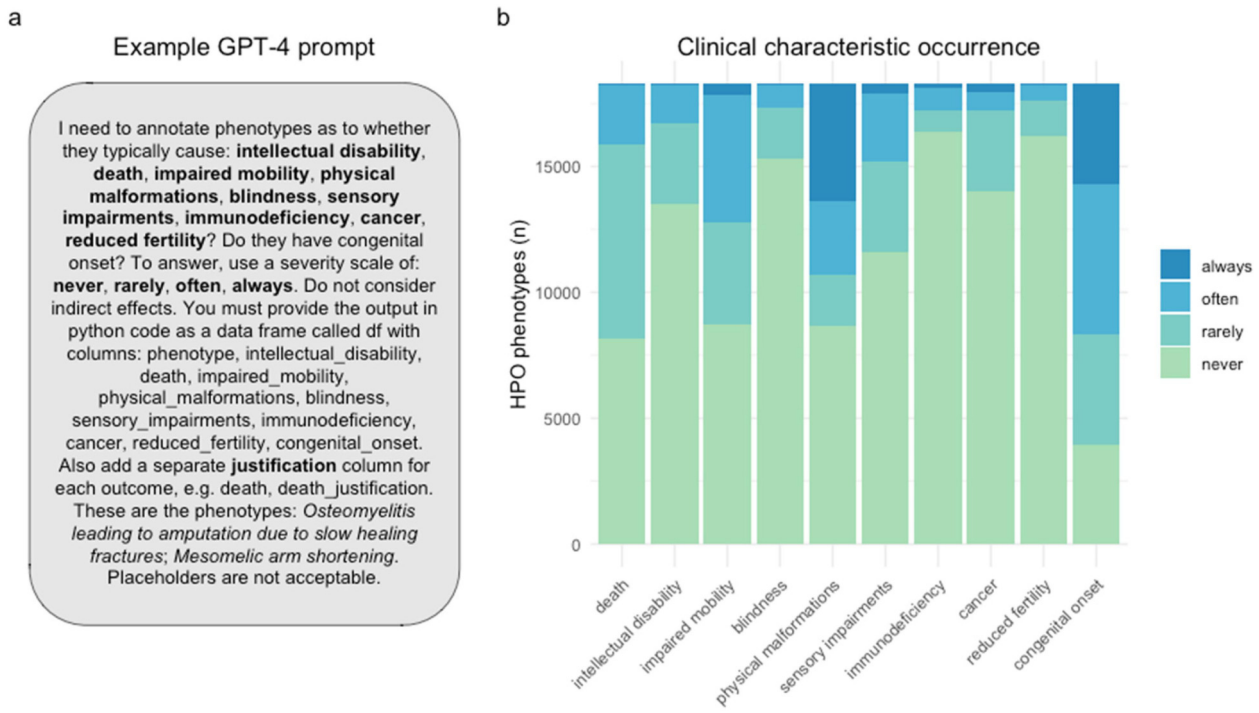

FIGURE 1
GPT-4 was able to annotate all human phenotypes based on whether they are always/often/rarely/never associated with different clinical characteristics. **(a)** An example of the prompt input given to to GPT-4. The phenotypes listed in the second to last sentence (*italicised*) were changed to allow all HPO phenotypes to be annotated. **(b)** Stacked bar plot showing the proportion of the occurrence of each clinical characteristic across all annotated HPO phenotypes. The terms shown on the *x*-axis are the clinical characteristics for which GPT-4 was asked to determine whether each phenotype caused them.

intellectual disability. While our annotations were not validated by human experts, this study was designed as a computational framework, with the primary aim of demonstrating what can be achieved at scale. Nevertheless, we acknowledge that LLM-generated annotations may carry systematic biases or inaccuracies. Future work could explore the incorporation of expert validation to further assess these and other outputs.

Quantifying phenotypic severity

While individual annotations are informative, we wanted to be able to distil the severity of each phenotype into a single score. Quantifying the overall severity of phenotypes can have important implications for diagnosis, prognosis, and treatment. It may also guide the prioritisation of gene therapy trials for phenotypes with the most severe clinical characteristics and thus the most urgent need. Importantly, the values reflected the severity of each clinical characteristic based on both the type of characteristic itself and its frequency within a particular phenotype. For instance, a phenotype always causing death would have a higher multiplied value than a phenotype often causing reduced fertility (see Table 1). First, we created a dictionary to map each clinical characteristic (e.g., blindness) and its frequency (always, often, rarely, never) to numeric values from 0 to 3. Then, the clinical characteristic values were multiplied by weights. Next, we computed an average score for each phenotype by aggregating the multiplied values across all clinical characteristics and then calculating the mean. This was then normalised by the theoretical maximum severity score, so that all phenotypes were on a 0–100 severity scale (where 100 is the most severe phenotype possible). This average normalised score represents the overall severity of the phenotype based on the severity of its individual clinical characteristics. Figure 3 illustrates the distribution of severity scores across HPO terms by branch, with "Abnormal cellular phenotype" being the branch that contains phenotypes with the highest severity scores on average.

Based on these scores we evaluated the top 50 severe phenotypes. One of the most severe phenotype was "Anencephaly" (HP:0002323) with a composite severity score of 45. Anencephaly is a birth defect where the baby is born without a portion of its brain and skull, often these babies are stillborn. In fact, many of the most severe phenotypes were related to developmental brain and neural tube defects. Comparison of the severity scores for each response, across the clinical characteristics annotated, revealed consistent trends: as the response of the clinical characteristic increased (from never to always), the severity score also increased (Supplementary Figure S10). We also evaluated the severity score distribution by HPO branch and calculated the mean severity score using all phenotypes within each major HPO branch (Figure 3). The HPO branch with the greatest mean severity score was "Abnormal cellular phenotype" (mean = 17), followed by "Neoplasm" (mean = 16.7), which would include the highly ranked phenotypes seen in Figure 4.

10.3389/fdgth.2026.1794934

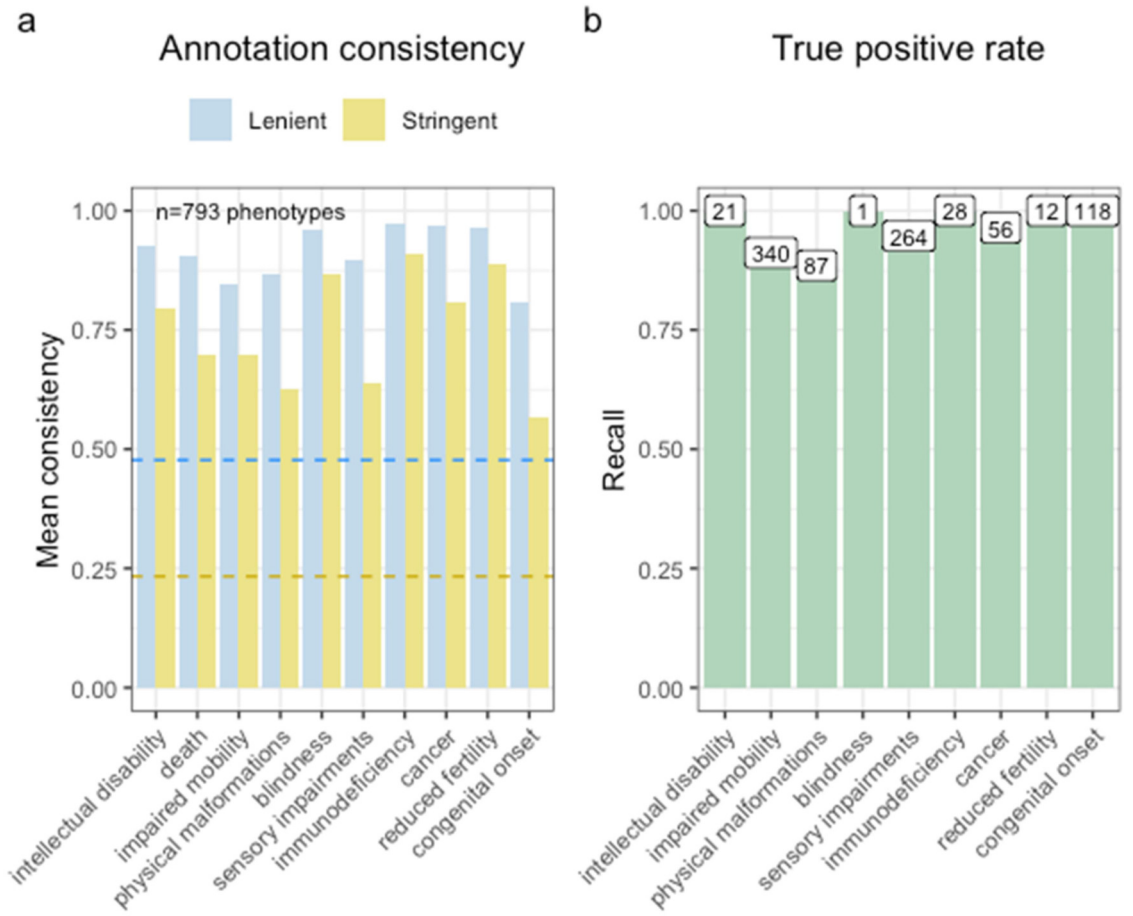

FIGURE 2
GPT-4 annotations are consistent and accurate across annotations. **(a)** Barplot showing the annotation consistency within phenotypes that were annotated more than once. In the lenient metric, annotations were collapsed into two groups ("always"/"often" and "never"/"rarely"). For a given clinical characteristic within a given phenotype, if an annotation was always within the same group it was considered consistent. In the stringent metric, all four annotation categories were considered to be different from one another. Thus, annotations were only defined as consistent if they were all identical. The blue dashed line indicates the probability of two annotations being consistent by chance in the lenient metric (~1/2). The gold dashed line is the probability of two annotations being consistent by chance in the stringent metric (~1/4). **(b)** Bar plot of the true positive rate for each annotation. The labels above each bar indicate the number of phenotypes tested.

TABLE 1 Weighted scores for each clinical characteristic and GPT-4 response category.

Clinical characteristic	Always (3)	Often (2)	Rarely (1)	Never (0)
Death (6)	18	12	6	0
Intellectual disability (5)	15	10	5	0
Impaired mobility (4)	12	8	4	0
Blindness (4)	12	8	4	0
Physical malformations (3)	9	6	3	0
Sensory impairments (3)	9	6	3	0
Immunodeficiency (3)	9	6	3	0
Cancer (3)	9	6	3	0
Reduced fertility (1)	3	2	1	0
Congenital onset (1)	3	2	1	0

10.3389/fdgth.2026.1794934

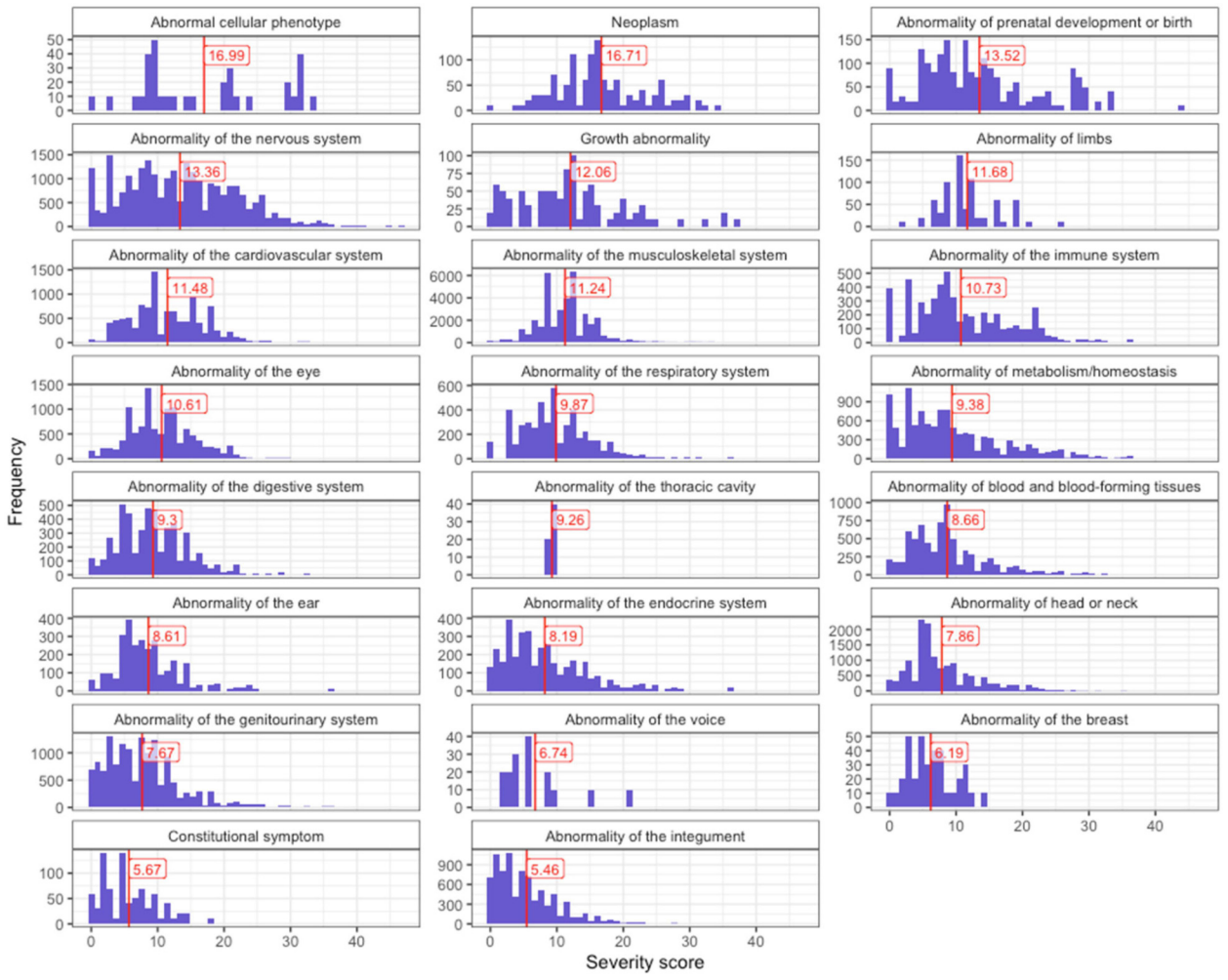

FIGURE 3
Distribution of composite GPT-4 severity scores across HPO terms, grouped by branch. The figure shows the distribution of phenotypes for each severity score (ranging from 0 to 100) across HPO branches. The vertical line indicates the average severity score for each branch. The branch "Abnormal cellular phenotype" contains the highest concentration of terms associated with the most severe clinical characteristics.

Severity classes

While the continuous severity score is a helpful metric, there may be some use cases where a categorical classification of severity is more immediately useful. In work by Lazarin et al. (17), the authors defined severity classed using a simple decision tree based on the individual severity annotations. We approximated this approach using our GPT-4 annotations. This categorical approach showed a strong degree of positive correspondence with the continuous severity score ($\widehat{\omega_p^2} = 0.88$, $p < 2.2 \times 10^{-308}$). In other words, severity score increased with severity class level (mild < moderate < severe < profound) as expected. The distribution of severity classes is shown in Supplementary Figure S12.

Correlations between clinical characteristic severity metrics

We found that some clinical characteristic severity metrics were correlated with one another, with a mean Pearson correlation of 0.2 across all individual metrics (see Supplementary Figure S11). In particular, blindness and sensory impairment were highly correlated with one another ($r = 0.62$, $p = 0$). Some metrics drove the composite severity score more than other, which is a reflection of both our per-metric weighting scheme, response type frequencies, and the correlation structure between metrics. Overall, impaired mobility seemed to be the strongest driver of the composite severity score with a Pearson correlation of 0.6001824, followed by intellectual disability ($r = 0.59$) and death ($r = 0.56$). This is driven by the greater number of phenotypes "always" causing impaired mobility (422), compared with, for instance, death (66), and intellectual disability (64).

Congenital onset by HPO branch

Next, we assessed the distribution of congenital onset across HPO branches (Figure 5). We found that the Abnormality of prenatal development or birth branch contained the greatest proportion of phenotypes that were always congenital (70.15%),

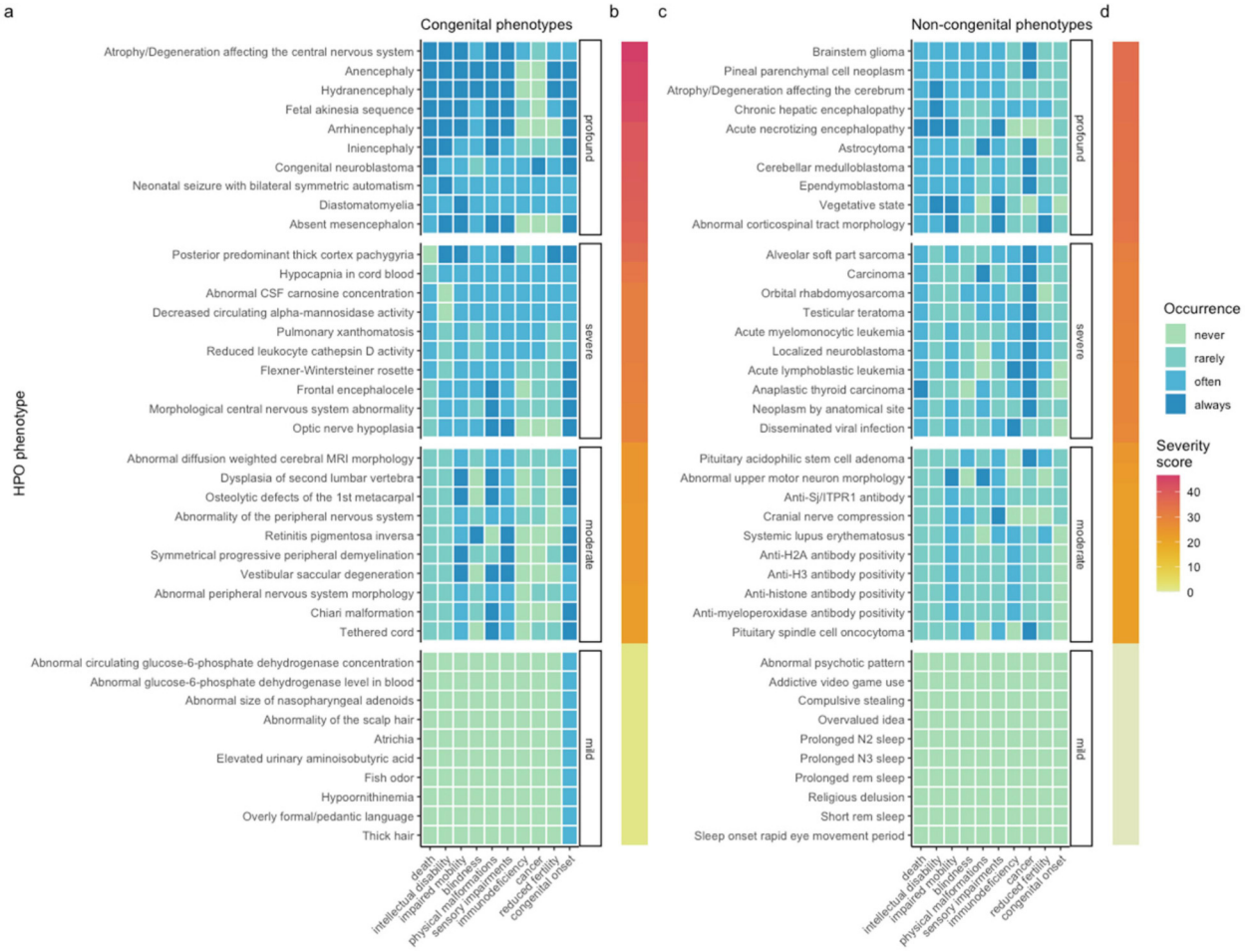

FIGURE 4
Quantifying the severity of HPO phenotype annotations highlights the most impactful conditions. Heatmap of 10 represetantive phenotypes from each severity class (Profound, Severe, Moderate, Mild) stratified by whether the phenotypes are often/always congenital **(a,b)** or rarely/never congenital **(c,d)**. Continuous severity scores are shown as bars **(b,d)** and were calculated by multiplying the numeric values assigned to each clinical characteristic according to Table 1. The average normalised score, representing overall phenotype severity on a 0-100 scale, was calculated by aggregating the multiplied values and normalising by the theoretical maximum severity score. The x-axes show each of the clinical characteristics. All data for this figure, as well as justifications for each annotation, can be found in Supplementary Table S3.

followed by Abnormality of prenatal development or birth (70.15%) and Abnormality of prenatal development or birth (70.15%). This is concordant with the expectation that these phenotypes should largely be congenital. The HPO branches with the least commonly congenital phenotypes were Constitutional symptom (0%), Constitutional symptom (0%), and Constitutional symptom (0%). "Constitutional symptom" is a fairly broad term defined as "*A symptom or manifestation indicating a systemic or general effect of a disease and that may affect the general well-being or status of an individual.*" Examples include "Fatigue" "Exercise intolerance", "Hot flashes" and "Sneeze".

Discussion

Phenotype severity annotations have utility across a wide variety of applications in both the clinic and research. In clinical settings, severity annotations can be used to prioritise the treatment of some phenotypes over others in patients with complex presentations, avoid administering contraindicated drugs, and prognosing potential health outcomes. In research settings, severity annotations can be used to identify phenotypes that have a large impact on patient outcomes and yet are currently understudied. They may also be used to help design new experiments and studies, or even provide new insights into the underlying aetiology of the disease by making expert-level summaries more immediately accessible to the wider research community.

The creation and annotation of biomedical knowledge has traditionally relied on manual or semi-manual curation by human experts (1, 2, 19–21). Performing such manual curation and review tasks at scale is often infeasible for human biomedical experts given limited time and resources. LLMs have the capacity to effectively encode, retrieve, and synthesise vast amounts of diverse information in a highly scalable manner (3, 8, 15). This makes them powerful tools that can be applied in a rapidly expanding variety of scenarios, including medical practice, research and data curation (8, 12, 22–24).

Here, we introduce a novel framework to leverage GPT-4 (15) to systematically annotate the severity of 17,502 phenotypic

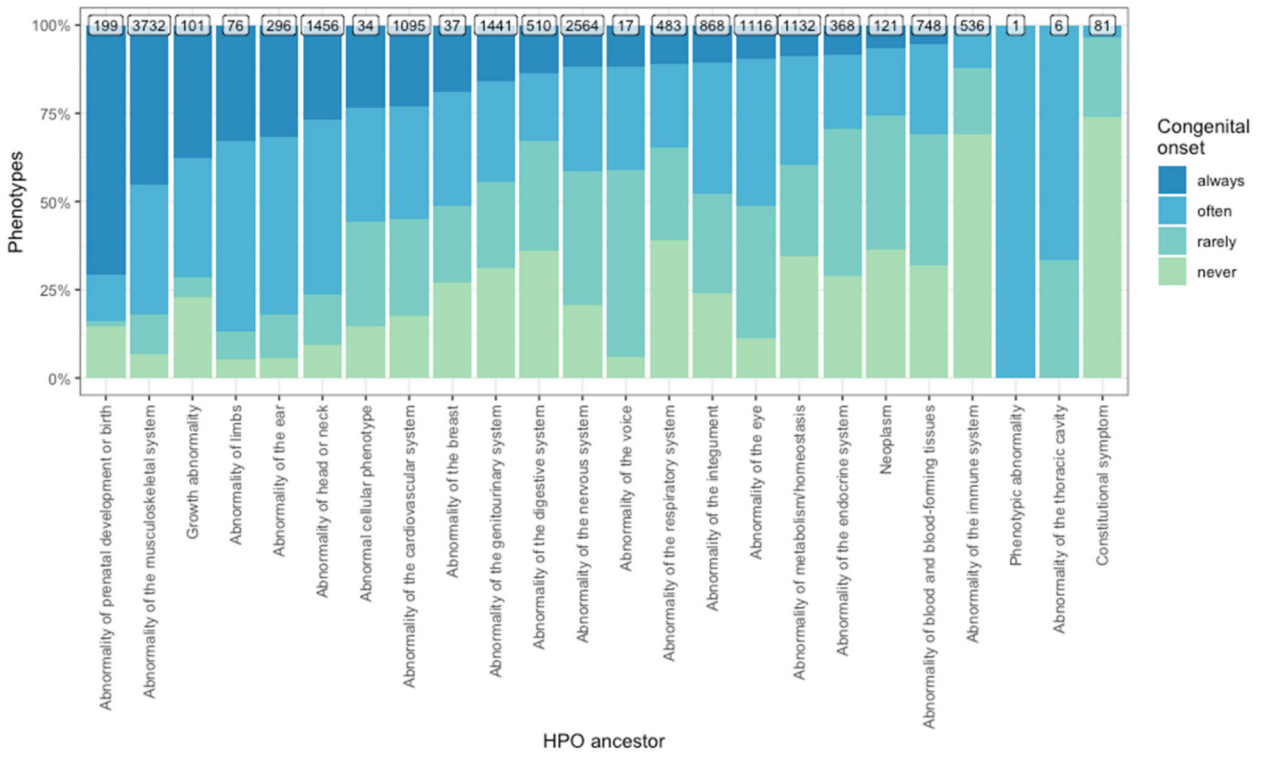

FIGURE 5
Distribution of congenital onset across HPO branches. The *y*-axis shows the proportion of phenotypes that are always/often/rarely/never congenital. The *x*-axis shows the HPO branch, orderered from highest to lowest proportion of always congenital phenotypes.

abnormalities within the HPO. By employing advanced AI capabilities, we have demonstrated the feasibility of automating this process, significantly enhancing efficiency without substantially compromising accuracy. Our validation approach yielded a high true positive rate exceeding 88% across the phenotypes tested. Furthermore, our approach can be readily adapted and scaled to accommodate the growing volume of phenotypic data. In total, the entire study cost $296.27 in queries to the OpenAI API. While we do not have a direct comparison, this likely represents a extremely small fraction of the total costs of such a study if performed manually by human experts charging at an hourly rate. Even if all human annotations were provided on a volunteer basis, this would still require hundreds if not thousands of hours of cumulative manual human labour. Using our approach, severity annotations for the entire HPO can be generated automatically at a rate of ~100 phenotypes/hour. Further optimisation of the annotation process and increased API rate limits could potentially accelerate this even further.

Throughout this study, we observed that GPT-4 was capable of reliably recovering deep semantic relationships from the medical domain, far beyond making superficial inferences based on lexical similarities. An excellent example of this is the phenotype "Molar tooth sign on MRI" (HP:0002419; severity score = 25.56), which GPT-4 annotated as causing intellectual disability. At first glance, we ourselves assumed this was a false positive as the term appeared to be related to dentition. However, upon further inspection we realised that molar tooth sign is in fact a pattern of abnormal brain morphology that happens to bear some resemblance to molar dentition when observed in radiological scans. This phenotype is a known sign of neurodevelopmental defects that can indeed cause severe intellectual disability (25).

In addition to rapidly synthesising and summarising vast amounts of information, LLMs can also be steered to provide justifications for each particular response. This makes LLMs amenable to direct interrogation as a means of recovering explainability, especially when designed to retain information about previous requests and interactions as they use these to iteratively improve and update their predictions (26). This represents a categorical advance over traditional natural language processing models based on more shallow forms of statistical or machine learning [e.g., Term Frequency-Inverse Document Frequency (27), Word2vec (28)] which lack the ability to provide chains of causal reasoning to justify their predictions. This highlights the fundamental trade-off between simpler models with high explainability (the ability humans to understand the inner workings of the model) but low interpretability (the ability of humans to trace the decision process of the model, analogous to human "reasoning"), and deeper more complex models with low explainability but high interpretability (29).

A key contribution of our study is the introduction of a quantitative severity scoring system that integrates both the nature of the clinical characteristic and the frequency of its occurrence. By encoding the concept of severity in this way, we are able to prioritise phenotypes based on their impact on patients. The methodology allowed us to transition from low-throughput qualitative assessments of severity [e.g., Lazarin et al.

(17)] to high-throughput quantitative assessments of severity. One of the most severe phenotypes in the HPO is "Fetal akinesia sequence" (FAS; HP:0001989, severity score = 43.9), and extremely rare condition that is almost always lethal. FAS is a complex, multi-system phenotype that can be caused by at least 24 different genetic disorders. Despite the complex and heterogeneous aetiology of this phenotype, GPT-4 was able to provide accurate annotations alongside explainable justifications for those annotations (see Supplementary Table S4). For example, this phenotype almost always results in death, either *in utero* or shortly after birth. Not only did GPT-4 correctly provide the annotation death as "always", when asked whether FAS causes sensory impairments it provided the response "always" with the justification "Fetal akinesia sequence typically results in severe sensory impairment due to neurodevelopmental disruption". Neurodevelopmental disruption is indeed a hallmark component of FAS (e.g., hydrocephalus, cerebellar hypoplasia) that causes severe impairments across multiple sensory systems (30). This demonstrates that GPT-4 was able to recover the correct chain of causality from phenotype to clinical characteristic.

Our findings highlight the potential of this next generation of natural language processing technologies in significantly contributing to the automation and refinement of data curation in biomedical research. These results have a large number of useful real-world applications, such as prioritising gene therapy candidates (31) and guiding clinical decision-making in rare diseases. It may also be used as tool to help inform policy decisions and funding allocation by healthcare or governmental institutions. This of course would need to be in consultation with subject matter medical experts, patients, advocates and biomedical ethicists before reaching a final decision. Nevertheless, access to succinct, interpretable, and semi-quantitative severity annotations may encourage key decision makers with limited time to review individual proposals to pay heed to phenotypes and diseases that would otherwise be overlooked. As the HPO and the broader literature continue to grow over time, our automated AI-based approach can easily be repeated to keep pace with the rapidly evolving biomedical landscape. Furthermore, it can be extended to produce different sets of annotations or be used with any other ontology. Additional use cases include gathering data on the prevalence of each phenotype to approximate their social and financial costs.

One limitation of this study is that it did not test how GPT-4 considers treatment availability when making annotations. Some severe conditions, like syphilis or certain melanomas, are curable with early detection and modern healthcare. GPT-4 often linked outcomes to access and quality of care, describing some cancers as fatal unless caught early while noting others are rarely lethal if treatment is available. This suggests the model accounts for healthcare quality but not consistently. The study also did not capture differences in severity within the same phenotype, such as light sensitivity ranging from mild to disabling. Nor did it consider onset, disease course, or how phenotypes change in specific contexts, such as hypertension being mild in most people but severe in rare genetic disorders. Other features relevant to severity, such as pain intensity and frequency, frequency of hospitalisations, lifetime burden, psychological functioning, and social functioning, were also not considered. A relevant example where such features would likely have been important in highlighting severity is endometriosis, defined as the growth of endometrial tissue outside the uterus. It affects approximately 10% of women of reproductive age and is associated with cyclical and chronic pelvic pain, infertility, and significant impairment of psychological and social functioning (32). It may also require repeated surgical intervention and hospitalisation. While our approach was able to capture severity associated with infertility, the associated pain burden, psychological and social impact, and potential need for surgery and hospitalisation were not reflected. These limitations extend beyond the annotations to the construction of the severity score itself, and highlight important areas for future work.

Therapeutic prioritization adds further challenges. Cost also shapes priorities, as some conditions require years of expensive care while others cause early death with little chance of intervention. Practical constraints must also be considered as long-term trials are difficult, whereas testing therapies that quickly reverse severe symptoms may yield faster results. These choices should involve ethicists, policymakers, advocacy groups, and families, with LLMs helping gather data at scale. While this study shows the promise of AI-driven phenotypic annotation, it also reveals risks of bias, gaps in severity scoring, and the need to factor in gene–environment interactions. Finally, a flurry of new LLMs continue to be released regularly. Here, we use GPT-4 as an example but note that our framework is equally applicable to most other LLMs. Future work should refine methods, add genomic and clinical data, and validate results to make AI a stronger tool in precision medicine.

In conclusion, our study represents a significant step towards harnessing the power of AI to advance phenotypic annotation and severity assessment across all rare diseases. This resource aims to provide researchers and clinicians with actionable insights that can inform rare disease research and improve the lives of individuals affected by rare diseases.

Methods

Annotating the HPO using OpenAI GPT-4

We wrote a Python script to iteratively query GPT-4 via the OpenAI application programming interface (API). The ultimately yielded consistently formatted annotations for 17,502 terms within the HPO. Our annotation framework was developed based on previously defined criteria for classifying disease severity (17). We sought to evaluate whether each phenotype directly caused a given severity-related clinical characteristic, including: intellectual disability, death, impaired mobility, physical malformations, blindness, sensory impairments, immunodeficiency, cancer, reduced fertility, and/or had a congenital onset. Through prompt engineering we found that the performance of GPT-4 improved when we incorporated a scale associated with each clinical characteristic and required a justification for each response. We asked how frequently the given phenotype directly causes each clinical characteristic—whether it never, rarely, often, or always occurred. This design helps to constrain the potential responses of GPT-4 and thus make it more amenable to machine-readable post-processing. It

also serves to address one of its key limitations from the Lazarin et al. (17) survey, namely the lack of information on how clinical characteristic frequency affected the clinicians' severity annotations. Here, we can instead use the frequency values to generate more precise annotations and downstream severity ranking scores.

Furthermore, our prompt design revealed that the optimal trade-off between the number of phenotypes and performance (in terms of producing the desired annotations, and adhering to the formatting requirements) was achieved when inputting no more than two or three phenotypes per prompt. An example prompt can be seen in Figure 1. Thus, only two phenotypes were included per prompt in order to (1) avoid exceeding per-query token limits, and (2) prevent the breakdown of GPT-4 performance due to long-form text input, which is presently a known limitation common to many LLMs including GPT-4 (33).

Calculating the true positive rate

A true positive rate was calculated as a measure of the recall of the GPT-4 annotations. This was achieved by identifying specific branches within the HPO that would contain phenotypes that would reliably indicate the occurrence of certain clinical characteristics, and using all descendants of this HPO branch as true positives. For example, all descendants of the terms "Intellectual disability" (HP:0001249) or "Mental deterioration" (HP:0001268) should be annotated as always or often causing intellectual disability (Table 2).

TABLE 2 The HPO branches and their descendants used as true positives for each clinical characteristic.

Clinical characteristic	HPO queries	True positive HPO IDs
Intellectual disability	"Intellectual disability"; "Mental deterioration"	19
Impaired mobility	"Gait disturbance"; "Diminished movement"; "mobility"	319
Physical malformations	"malformation"	78
Blindness	"blindness"	1
Sensory impairments	"Abnormality of vision"; "Abnormality of the sense of smell"; "Abnormality of taste sensation"; "Somatic sensory dysfunction"; "Hearing abnormality"	252
Immunodeficiency	"Immunodeficiency"; "Impaired antigen-specific response"	29
Cancer	"Cancer"; "malignant"; "carcinoma"	56
Reduced fertility	"Decreased fertility"; "Hypogonadism"	9

Quantifying phenotypic severity

The GPT-4 generated clinical characteristic occurrences were converted into a semi-quantitative scoring system, with "always" corresponding to 3, "often" to 2, "rarely" to 1, and "never" to 0. These scores were then weighted by a severity metric on a scale of 1–6, with 6 representing the highest severity. The weights were based on the severity of each clinical characteristic (Table 1), using the "Tiers" defined in Lazarin et al. (17) (shorted life span in infancy/childhood/adolescence = 1, intellectual disability = 1, shortened lifespan in adulthood = 2, impaired mobility = 2, physical malformation = 2, sensory impairments = 3, immunodeficiency = 3, cancer = 3, and reduced fertility = 4). Note, that the scale in Lazarin et al. (17) is reversed, with 1 being the most severe. In their study, the tiers were based on rankings on a scale of 1–10 in terms of how important each disease clinical characteristic was to determine disease severity. An average ranking of >9 placed a clinical characteristic in Tier 1, 8 in Tier 2, 6–7 in Tier 3 and below 6 in Tier 4.

Let us denote:

- p: a phenotype in the HPO.
- j: the identity of a given annotation metric (i.e., clinical characteristic, such as "intellectual disability" or "congenital onset").
- W_j: the assigned weight of metric j.
- F_j: the maximum possible value for metric j (equivalent across all j).
- F_{pj}: the numerically encoded value of annotation metric j for phenotype p.
- NSS_p: the final composite severity score for phenotype p after applying normalisation to align values to a 0–100 scale and ensure equivalent meaning regardless of which other phenotypes are being analysed in addition to p. This allows for direct comparability of severity scores across studies with different sets of phenotypes.

Severity classes

The decision tree algorithm used in Lazarin et al. (17) was adapted here for use with the GPT-4 clinical characteristic annotations. This algorithm first assigned each clinical chacteristic to a tier, where Tier 1 indicated the most severe clinical characteristics and Tier 4 indicated the least severe clinical characteristics ("death" = 1, "intellectual disability" = 1, "impaired mobility" = 2, "physical malformations" = 2, "blindness" = 3, "sensory impairments" = 3, "immunodeficiency" = 3, "cancer" = 3, "reduced fertility" = 4). If a phenotype often or always caused more than one Tier 1 clinical characteristic, it was assigned a severity class of "Profound". If the phenotype often or always caused only one Tier 1 clinical characteristic, it was assigned a severity class of "Severe". A "Severe" class assignment was also assigned if the phenotype often or always caused three or more Tier 2 and Tier3 clinical characteristics. If the phenotype often or always caused at least one Tier 2 clinical characteristic, it was assigned a severity class of "Moderate". All remaining phenotypes were was assigned a severity class of "Mild". In cases where the

phenotype mapped to more than one class, only the most severe class was used. This procedure is implemented within the function **HPOExplorer::gpt_annot_class**.

Correlations between clinical characteristic severity metrics

To assess the correlation structure between each clinical characteristic severity metric, as well as between the composite severity score and each metric, we computed Pearson correlation coefficients for all pairwise combinations of these variables using the numerically encoded metric values. The correlation matrix was visualised using a heatmap, with the colour intensity representing the strength of the correlation (Supplementary Figure S11).

Data availability statement

All code and data used in this study are available on Github at: https://github.com/neurogenomics/gpt_hpo_annotations. The GPT-4 clinical characteristic annotations for all HPO phenotypes are made available through the R function HPOExplorer::gpt_annot_read or in CSV format at: https://github.com/neurogenomics/gpt_hpo_annotations/tree/master/data. A fully reproducible version of this Quarto manuscript can be found at: https://github.com/neurogenomics/gpt_hpo_annotations/blob/master. Further inquiries can be directed to the corresponding author/s.

Author contributions

KM: Conceptualization, Writing – review & editing, Funding acquisition, Validation, Project administration, Formal analysis, Supervision, Methodology, Data curation, Investigation, Software, Visualization, Writing – original draft, Resources. BS: Writing – review & editing, Formal analysis, Resources, Supervision, Methodology, Project administration, Writing – original draft, Software, Data curation, Visualization, Investigation, Funding acquisition, Validation, Conceptualization. NS: Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by a UK Dementia Research Institute (UK DRI) Future Leaders Fellowship (MR/T04327X/1) and the UK DRI which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

Acknowledgments

We would like to thank members of the Monarch Initiative for their insight and feedback throughout this project. In particular, Peter Robinson.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1794934/full#supplementary-material

References

1. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. *Nucleic Acids Res.* (2021) 49(D1):D1207–17. doi: 10.1093/nar/gkaa1043

2. Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, et al. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* (2024) 52(D1):D1333–46. doi: 10.1093/nar/gkad1005

3. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med.* (2024) 30:1134–42. doi: 10.1038/s41591-024-02855-5

4. Bolton E, Venigalla A, Yasunaga M, Hall D, Xiong B, Lee T, et al. BioMedLM: a 2.7B parameter language model trained on biomedical text. *arXiv.* (2024). doi: 10.48550/arXiv.2403.18421

5. Zhang K, Yu J, Yan Z, Liu Y, Adhikarla E, Fu S, et al. BiomedGPT: a unified and generalist biomedical generative Pre-trained transformer for vision, language, and multimodal tasks. *arXiv.* (2023). doi: 10.48550/arXiv.2305.17100

6. Labrak Y, Bazoge A, Morin E, Gourraud P-A, Rouvier M, Dufour R. Biomistral: a collection of open-source pretrained large language models for medical domains. *arXiv.* (2024). doi: 10.48550/arXiv.2402.10373

7. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthcare*. (2021) 3(1):1–23. doi: 10.1145/3458754

8. Singhal K, Azizi S, Tu T, Sara Mahdavi S, Wei J, Chung HW, et al. 'Large language models encode clinical knowledge. *Nature*. (2023) 620:172–80. doi: 10.1038/s41586-023-06291-2

9. Singhal K, Tu T, Gottweis J, Sayres R, Ellery Wulczyn LH, Clark K, et al. Towards expert-level medical question answering with large language models. *arXiv*. (2023b). doi: 10.48550/arXiv.2305.09617

10. Shin H-C, Zhang Y, Bakhturina E, Puri R, Patwary M, Shoeybi M, et al. Biomegatron: larger biomedical domain language model. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2020). p. 4700–6. doi: 10.18653/v1/2020.emnlp-main.379

11. Cheng K, Guo Q, He Y, Lu Y, Gu S, Wu H. Exploring the potential of GPT-4 in biomedical engineering: the dawn of a new era. *Ann Biomed Eng*. (2023) 51(8):1645–53. doi: 10.1007/s10439-023-03221-1

12. O'Neil ST, Schaper K, Elsarboukh G, Reese JT, Moxon SAT, Harris NL, et al. Phenomics assistant: an interface for LLM-based biomedical knowledge graph exploration. *bioRxiv*. (2024). doi: 10.1101/2024.01.31.578275

13. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative Pre-trained transformer for biomedical text generation and mining. *Brief Bioinformatics*. (2022) 23(6):bbac409. doi: 10.1093/bib/bbac409

14. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. *arXiv*. (2023). doi: 10.48550/arXiv.2312.00164

15. OpenAI JA, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. *arXiv*. (2024). doi: 10.48550/arXiv.2303.08774

16. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv*. (2023). doi: 10.48550/arXiv.2311.16452

17. Lazarin GA, Hawthorne F, Collins NS, Platt EA, Evans EA, Haque IS. Systematic classification of disease severity for evaluation of expanded carrier screening panels. *PLoS One*. (2014) 9(12):e114391. doi: 10.1371/journal.pone.0114391

18. Garagnani L, Smith GD. Syndromes associated with syndactyly. In: Abzug JM, Kozin S, Zlotolow DA, editors. *The Pediatric Upper Extremity*. New York, NY: Springer (2013). p. 1–31. doi: 10.1007/978-1-4614-8758-6_14-1

19. Putman TE, Schaper K, Matentzoglu N, Rubinetti VP, Alquaddoomi FS, Cox C, et al. The monarch initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Res*. (2024) 52(D1):D938–49. doi: 10.1093/nar/gkad1082

20. Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res*. (2021) 49(D1):D1302–10. doi: 10.1093/nar/gkaa1027

21. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*. (2017) 45(D1):D712–22. doi: 10.1093/nar/gkw1128

22. Toro S, Anagnostopoulos AV, Bello S, Blumberg K, Cameron R, Carmody L, et al. Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI). *arXiv*. (2023). doi: 10.48550/arXiv.2312.10904

23. Pan JZ, Razniewski S, Kalo J-C, Singhania S, Chen J, Dietze S, et al. Large language models and knowledge graphs: opportunities and challenges. *arXiv*. (2023). doi: 10.48550/arXiv.2308.06374

24. Caufield JH, Hegde H, Emonet V, Harris NL, Joachimiak MP, Matentzoglu N, et al. Structured prompt interrogation and recursive extraction of semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. *arXiv*. (2023). doi: 10.48550/arXiv.2304.02711

25. Gleeson JG, Keeler LC, Parisi MA, Marsh SE, Chance PF, Glass IA, et al. Molar tooth sign of the midbrain–hindbrain junction: occurrence in multiple distinct syndromes. *Am J Med Genet Part A*. (2004) 125A(2):125–34. doi: 10.1002/ajmg.a.20437

26. Janik RA. Aspects of human memory and large language models. *arXiv*. (2024). doi: 10.48550/arXiv.2311.03839

27. Jones KS. A statistical interpretation of term specificity and its application in retrieval. *J Doc*. (1972) 28(1):11–21. doi: 10.1108/eb026526

28. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv*. (2013). doi: 10.48550/arXiv.1301.3781

29. Marcinkevičs R, Vogt JE. Interpretability and explainability: a machine learning zoo mini-tour. *arXiv*. (2023). doi: 10.48550/arXiv.2012.01805

30. Chen C-P. Prenatal diagnosis and genetic analysis of fetal akinesia deformation sequence and multiple pterygium syndrome associated with neuromuscular junction disorders: a review. *Taiwan J Obstet Gynecol*. (2012) 51(1):12–7. doi: 10.1016/j.tjog.2012.01.004

31. Schilder BM, Murphy KB, Dash H, Zhang Y, Gordon-Smith R, Chapman J, et al. Cell type-specific contextualisation of the human phenome: towards the systematic treatment of all rare diseases. *medRxiv*. (2025). doi: 10.1101/2023.02.13.23285820

32. Zondervan KT, Becker CM, Koga K, Missmer SA, Taylor RN, Viganò P. Endometriosis. *Nat Rev Dis Primers*. (2018) 4(1):9. doi: 10.1038/s41572-018-0008-5

33. Wei J, Yang C, Song X, Lu Y, Hu N, Huang J, et al. Long-form factuality in large language models. *arXiv*. (2024) doi: 10.48550/arXiv.2403.18802