Note-Level Phenotyping of Multiple Sclerosis Notes by a Large Language Model Achieves Near Human-Level Agreement

Daniel B. Hier *, Pavankumar Y. Srinivasula, Michael D. Carrithers

Posted Date: 16 April 2026

doi: 10.20944/preprints202604.1173.v1

Keywords: multiple sclerosis; phenotype; clinical text; large language models; natural language processing; inter-rater agreement; phenotyping

 doi:10.20944/preprints202604.1173.v1

Article

Note-Level Phenotyping of Multiple Sclerosis Notes by a Large Language Model Achieves Near Human-Level Agreement

Daniel B. Hier [1,*], Pavankumar Y. Srinivasula [2], and Michael D. Carrithers [1]

[1] College of Medicine, University of Illinois at Chicago, Chicago, IL 60612, USA
[2] College of Liberal Arts and Sciences, University of Illinois at Chicago, Chicago, IL 60607, USA
* Correspondence: dhier@uic.edu

Abstract

Background/Objectives: Clinical phenotyping from narrative electronic health records (EHRs) often relies on multi-stage pipelines with span-level extraction, ontology mapping, and aggregation, which are complex to develop and maintain. Large language models (LLMs) may enable direct note-level abstraction of clinically meaningful features without intermediate extraction steps. We evaluated whether an LLM can approximate human inter-rater agreement for note-level multiple sclerosis (MS) phenotyping. **Methods:** We analyzed 100 de-identified MS neurology progress notes from a single academic medical center, each annotated for the presence or absence of 17 predefined neurological phenotype features (e.g., weakness, sensory, gait, pain, cognition, bladder). Two human annotators independently labeled all notes using a multi-label note-level framework in Prodigy, and discrepancies between the human annotators were adjudicated to create a gold standard. The same notes were annotated in a zero-shot setting by a large language model (GPT-5) and by the dictionary-based Doc2Hpo system. We computed percent agreement, Cohen's κ, and precision, recall, and F1 scores for each annotator relative to the gold standard **Results:** Human–human agreement was substantial across most phenotype domains, with Cohen's κ typically between 0.61 and 0.84, and lower agreement for infrequent features such as spasticity and hyperreflexia. Agreement between the LLM and human annotators was comparable to human inter-rater agreement across many features. Relative to the gold standard, the LLM showed recall that was equal to or higher than human annotators for most phenotypes, with overall F1 scores similar to human performance, whereas Doc2Hpo demonstrated lower precision and recall. **Conclusions:** In this note-level MS phenotyping task, a large language model achieved performance approaching expert human inter-rater agreement, with particularly strong recall across multiple phenotype domains. These findings suggest that direct note-level abstraction of clinically meaningful phenotypes from narrative neurology notes is feasible and may offer a scalable complement to traditional span-oriented extract–map–aggregate pipelines for population-level phenotyping and downstream machine learning applications.

Keywords: multiple sclerosis; phenotype; clinical text; large language models; natural language processing; inter-rater agreement; phenotyping

1. Introduction

The phenotype of multiple sclerosis (MS) comprises its disease course [1,2] and clinical presentation [3]. The clinical presentation phenotype is the pattern of neurologic signs and symptoms in an individual patient, including motor, sensory, visual, cerebellar, brainstem, sphincter, and cognitive manifestations [3,4]. In research settings, such features may be represented as structured variables, such as the components of the Functional Systems Scale (FSS) of the Kurtzke Expanded Disability Status Scale (EDSS) [5]. In routine clinical care, however, clinical presentation phenotype is often

 doi:10.20944/preprints202604.1173.v1

documented in free-text neurology notes, creating challenges for systematic extraction and comparison across patients [6].

Extraction of clinical presentation phenotype from the unstructured free text in electronic health records (EHRs) is important for MS research and may also support clinical care [6]. In both clinical trials and routine practice, assessment of symptom burden is central to evaluating response to disease-modifying therapy (DMT) and informing therapeutic decision-making [7,8]. Swetlik et al. [9] recommended including standardized discrete data elements in each clinical note (e.g., EDSS, relapse status, disease course, initial presentation, and DMT), but such documentation has not been widely adopted, and substantial clinically relevant information remains embedded in free-text narrative.

Automated extraction of phenotypes from EHRs has emerged as a central problem in biomedical informatics, owing to the large proportion of clinically relevant information that resides in unstructured narrative text rather than standardized discrete fields [10,11]. Such phenotypic data support cohort discovery, longitudinal modeling, and secondary use of clinical data for observational and translational research. In the MS domain, prior work has demonstrated that detailed disease traits, including clinical subtype and disability measures, can be extracted from routine electronic records with high reliability [6]. However, precise phenotype characterization in most settings still relies heavily on manual review of clinician notes, which is labor-intensive and difficult to scale.

Prior work has demonstrated that high levels of inter-rater agreement can be achieved for neurologic concept annotation when structured annotation tools such as Prodigy are used with clear guidelines [12]. In that study, inter-rater reliability was assessed for text-span identification of neurological signs and symptoms within EHR notes, and agreement between human annotators and a convolutional neural network (CNN) was evaluated. Although human–human agreement was substantial, machine–human agreement remained lower than human–human agreement.

Traditional approaches to phenotype extraction from clinical text have relied on supervised machine learning or rule-based systems that require annotated training corpora and domain-specific feature engineering. While effective for narrowly defined tasks, these approaches may be limited in their ability to capture the full complexity of neurologic phenotypes, which often require integration of distributed contextual information across a clinical note. In contrast, LLMs offer the ability to perform zero-shot inference by leveraging extensive pretraining on biomedical and general text corpora, enabling them to interpret clinical narratives without task-specific fine-tuning. This capability is particularly relevant for MS phenotype identification, where clinical presentation is heterogeneous and often described implicitly rather than through standardized terminology. In this framework, roll-up or subsumption is no longer an explicit post-processing step but becomes an intrinsic property of the model's interpretation. Accordingly, zero-shot LLM-based approaches may provide a scalable alternative to traditional methods for extracting clinically meaningful phenotypes from unstructured narrative documentation.

Because clinical interpretation of narrative documentation inherently involves subjective judgment, human inter-rater reliability provides a critical benchmark for evaluating automated systems. Cohen's κ is widely used to quantify agreement beyond chance and serves as a standard measure in clinical annotation studies [13,14]. Establishing the level of agreement achievable between trained human annotators for multiple sclerosis phenotype features is essential for determining whether automated approaches attain useful performance. Perfect concordance in clinical note annotation is unlikely for either human-human or human-machine comparisons. Nonetheless, the concordance between human annotators can define a useful upper bound for annotation concordance between machine and human annotators. The goal of machine annotation systems is therefore to approach the level of human-human agreement rather than to exceed it.

Recent advances in large language models (LLMs) have demonstrated strong performance across a range of medical natural language processing tasks without task-specific fine-tuning [15–18]. Recent work applying LLMs to high-throughput phenotyping of physician notes has shown that LLM-based approaches can outperform traditional deep learning and classical machine learning methods [19].

 doi:10.20944/preprints202604.1173.v1

These models may enable scalable extraction of phenotypes and disease status directly from narrative EHR documentation in zero-shot settings.

In this study, phenotype identification is performed at a higher level of abstraction, focusing on the presence or absence of clinically meaningful phenotype categories rather than fine-grained concept extraction or normalization to controlled vocabularies such as SNOMED CT or the Human Phenotype Ontology (HPO). Although ontology linking is an important component of many biomedical NLP pipelines, it represents a distinct task that introduces additional sources of variability and error. Prior work has shown that LLMs may accurately interpret clinical concepts while failing to reliably map those concepts to standardized identifiers. By restricting the task to high-level phenotype classification, the present study isolates the problem of clinical interpretation from the downstream challenge of ontology normalization, allowing a more direct comparison between human and machine understanding of narrative clinical text. In this context, phenotype is considered in two complementary senses: disease course classification and clinical presentation based on neurologic signs and symptoms.

The central question addressed in this study is whether a large language model can read a neurology note and arrive at the same clinical interpretation of these phenotypes as a human observer, as reflected by inter-rater agreement metrics. Large language models can approximate clinician-derived abstraction of high-level disease phenotypes from narrative clinical text.

2. Methods

2.1. Neurology Notes

A total of 12,661 de-identified neurology clinical notes from the University of Illinois Hospital, the primary teaching hospital of UI Health, were obtained from a REDCap database for the period January 14, 2016, through September 2, 2022. Notes were filtered to include only those longer than 600 words, associated with a diagnosis of multiple sclerosis (ICD-10-CM G35), and classified as "Progress Note" encounters. After deduplication, 4,617 notes remained for analysis.

Each clinical note represents a narrative summary of a patient encounter and served as the unit of analysis in this study. Prior to annotation, notes were converted to JavaScript Object Notation (JSON) format, with each complete note represented as a single JSON object. This note-level representation preserves the full clinical context of the note, enabling evaluation of disease status and phenotype feature identification as a task of clinical interpretation rather than isolated text-span extraction.

Use of the de-identified clinical documentation for research was approved by the Institutional Review Board of the University of Illinois (Protocol No. 2017-0520Z).

2.2. Human Annotation Process

Human annotation was performed using `Prodigy` (Explosion AI, Berlin, Germany), a commercial annotation platform for natural language processing workflows. Prodigy provides a locally hosted web interface and integrates with the `spaCy` library. Each clinical note was presented as a single annotation unit.

This study was designed as an inter-rater agreement experiment on a subset of 100 notes, randomly sampled from 4,617 eligible MS progress notes. The objective was to estimate human–human and human–LLM agreement under controlled conditions rather than to perform large-scale annotation.

Two annotators independently labeled the notes: Annotator 1 (A1), a pre-medical student, and Annotator 2 (A2), a senior neurologist, enabling assessment across differing levels of clinical expertise.

Written definitions of phenotype features (Appendix C) were provided prior to annotation, and joint training sessions using representative notes were conducted to promote consistency. Disease diagnoses (e.g., multiple sclerosis, optic neuritis) were not annotated, and modifiers such as laterality, severity, and duration were not separately coded.

Annotation was performed using Prodigy's `textcat.manual` recipe. Phenotype features (e.g., ataxia, weakness, tremor) were annotated using a multi-label classification interface, allowing multiple non-exclusive categories to be assigned to each note.

 doi:10.20944/preprints202604.1173.v1

Annotations were stored in an SQLite database and exported in JSON format for analysis.

The annotation framework operated at the level of the complete clinical note. Phenotypes were recorded as present or absent, without counting repeated mentions, and span-based methods such as named entity recognition were not used.

2.3. Annotation by Doc2Hpo

To establish a baseline for annotation performance, we applied the span-detecting dictionary-based concept extraction tool *Doc2Hpo* using its standard API configuration (https://doc2hpo.wglab.org/parse/acdat). This approach represents a traditional span-based, dictionary-driven paradigm relying on exact string matching rather than contextual interpretation. The interface accepts free-text input and returns recognized Human Phenotype Ontology (HPO) terms along with their corresponding ontology identifiers [20]. Text was processed with default settings, employing the Aho–Corasick string-matching algorithm against all HPO terms and synonyms within the Phenotypic Abnormality subtree (HP:0000118), with negation detection enabled, and retaining only the longest match when overlapping annotations were identified. Output includes length and start position of each text span detected.

For example, the input text:

```
The patient presented with multiple sclerosis with ataxia, increased reflexes,
paraparesis, and optic neuritis.
```

produced the following output, including negation status and character span information:

```
HP:0001251 | ataxia | negated=False | start=61 | length=6
HP:0001347 | increased reflexes | negated=False | start=69 | length=18
HP:0002385 | paraparesis | negated=False | start=89 | length=11
HP:0100653 | optic neuritis | negated=False | start=106 | length=14
```

2.4. Annotation by LLM

The same set of 100 clinical notes used for human annotation was independently evaluated using a state-of-the-art large language model, GPT-5 (OpenAI; `gpt-5.2-2025-12-11`), accessed via the OpenAI Responses API; in the remainder of this manuscript, we refer to this model as "the LLM." The temperature was set to 0 to ensure deterministic outputs, and no explicit maximum token limit was specified, with default model settings used for response length. The task was framed as multi-label classification at the level of the complete clinical note, using the same phenotype definitions provided to human annotators. Each note was processed independently using a fixed prompt. Model outputs were constrained using a predefined JSON schema to ensure structured and reproducible annotations. For each note, the model returned a list of phenotype items, each containing:

```
- label: phenotype category (from predefined list)
- present: boolean indicating presence or absence
- evidence: supporting text excerpt from the note
- detail: optional clarification
```

An optional warning field was included to capture uncertainty or ambiguity in the model output. The full prompt, label definitions, and implementation details are provided in Appendices A, B, and C.

2.5. Adjudication Process to Create the Gold Standard Annotation Set

A gold standard annotation set was constructed through a structured adjudication process to resolve discrepancies between annotators. The dataset comprised 100 clinical notes and 17 phenotypic features, yielding 1,700 note–feature evaluations. For 1,535 (90.3%) note–feature pairs, both human annotators (A1 and A2) were in agreement, and the gold standard label was assigned directly based on this consensus. In 165 (9.7%) cases, the human annotators were discordant; these cases were re-reviewed and adjudicated by the senior neurologist (A2). Adjudication resulted in agreement with A2

 doi:10.20944/preprints202604.1173.v1

in 93 cases and with A1 in 72 cases. This process yielded a fully adjudicated gold standard annotation set comprising 1,700 note–feature labels.

2.6. Statistical Analysis

Human–human and human–LLM inter-rater agreement was assessed using unadjusted agreement (percent concordance) and Cohen's κ, a chance-corrected measure of agreement [13,14]. Each annotation method (A1, A2, LLM, and Doc2Hpo) was compared to the gold standard annotation set to calculate precision, recall, and F1 scores for each phenotype using standard definitions. All analyses were performed using Python (including `openai`, `pandas`, `matplotlib`, `sklearn.metrics`, `json`, `collections`, and `requests` libraries).

3. Results

We examined agreement between two human annotators on 100 electronic health record (EHR) notes from patients with multiple sclerosis. Each note was annotated for the presence or absence of seventeen predefined neurological phenotype features, including `weakness`, `sensory`, `cognitive`, `gait`, `pain`, and `falls`. Raw (unadjusted) agreement between the human annotators was generally high and adjusted agreement (κ) was high (> 0.50) except for some features (dysphagia, ataxia, bowel, spasticity, and hyperreflexia (Table 1).

Table 1. Concordance Between Annotators on phenotype features (A1 vs. A2).

Phenotype	Agreement	κ	A1 positive	A2 positive
Falls	0.97	0.84	11	10
Tremor	0.98	0.74	4	4
Dysarthria	0.96	0.73	9	7
Vision	0.88	0.70	28	28
Psych	0.92	0.67	15	13
Fatigue	0.94	0.67	11	9
Pain	0.87	0.67	26	27
Gait	0.82	0.61	40	30
Cognitive	0.93	0.60	12	7
Sensory	0.83	0.59	28	31
Weakness	0.79	0.56	40	41
Bladder	0.91	0.52	13	8
Dysphagia	0.96	0.48	3	5
Ataxia	0.88	0.45	17	7
Bowel	0.93	0.33	7	4
Spasticity	0.84	0.13	14	6
Hyperreflexia	0.94	-0.02	1	5

The LLM was evaluated for the ability to annotate the same notes in a zero-shot prompt mode (Appendix B). We did not do any fine-tuning or retrieval-augmented generation. Agreement between the LLM and each human annotator was comparable to human inter-rater agreement (Tables 2 and 3).

Based on the annotations by the two human annotators, we created an adjudicated gold standard annotation set. For 1,535 of the 1,700 possible note–feature evaluations (17 features × 100 notes), the two annotators were in agreement. The remaining 165 discordant evaluations were adjudicated by the senior neurologist (A2), with 72 resolved in favor of A1 and 93 in favor of A2. Using this gold standard, we calculated precision, recall, and F1 for A1, A2, and the LLM. As a comparator, we also evaluated the text-span-identifying dictionary-based *Doc2Hpo* annotator. The LLM demonstrated consistently high recall across phenotypes, often exceeding that of human annotators, while maintaining comparable overall F1 scores. In contrast, Doc2Hpo demonstrated lower performance across both precision and recall relative to the LLM and human annotators (Figures 1, 2, and 3).

 doi:10.20944/preprints202604.1173.v1

Table 2. Concordance Between Annotators on phenotype features (A1 vs. LLM).

Phenotype	Agreement	κ	A1 positive	LLM positive
Dysarthria	0.98	0.86	9	7
Fatigue	0.97	0.86	11	14
Cognitive	0.96	0.82	12	14
Falls	0.94	0.72	11	13
Ataxia	0.92	0.70	17	15
Weakness	0.85	0.69	40	41
Gait	0.84	0.68	40	46
Vision	0.85	0.65	28	35
Tremor	0.96	0.65	4	8
Pain	0.82	0.61	26	42
Bowel	0.93	0.60	7	12
Dysphagia	0.96	0.58	3	7
Bladder	0.89	0.56	13	16
Psych	0.84	0.55	15	29
Sensory	0.79	0.52	28	35
Spasticity	0.83	0.03	14	5
Hyperreflexia	0.93	-0.02	1	6

Figure 1. Phenotype-level recall scores for each annotation method relative to the gold standard. Values across 17 phenotypic features highlight differences in method-specific performance and variability across clinical domains.

 doi:10.20944/preprints202604.1173.v1

Table 3. Concordance Between Annotators on phenotype features (A2 vs. LLM).

Phenotype	Agreement	κ	A2 positive	LLM positive
Hyperreflexia	0.99	0.90	5	6
Dysarthria	0.98	0.85	7	7
Sensory	0.90	0.77	31	35
Fatigue	0.95	0.76	9	14
Weakness	0.88	0.75	41	41
Vision	0.89	0.75	28	35
Falls	0.93	0.66	10	13
Tremor	0.96	0.65	4	8
Pain	0.83	0.63	27	42
Cognitive	0.93	0.63	7	14
Gait	0.82	0.63	30	46
Bladder	0.92	0.63	8	16
Psych	0.84	0.54	13	29
Spasticity	0.95	0.52	6	5
Ataxia	0.90	0.50	7	15
Dysphagia	0.94	0.47	5	7
Bowel	0.92	0.47	4	12

Figure 2. Phenotype-level precision scores for each annotation method relative to the gold standard. Values across 17 phenotypic features highlight differences in method-specific performance and variability across clinical domains.

 doi:10.20944/preprints202604.1173.v1

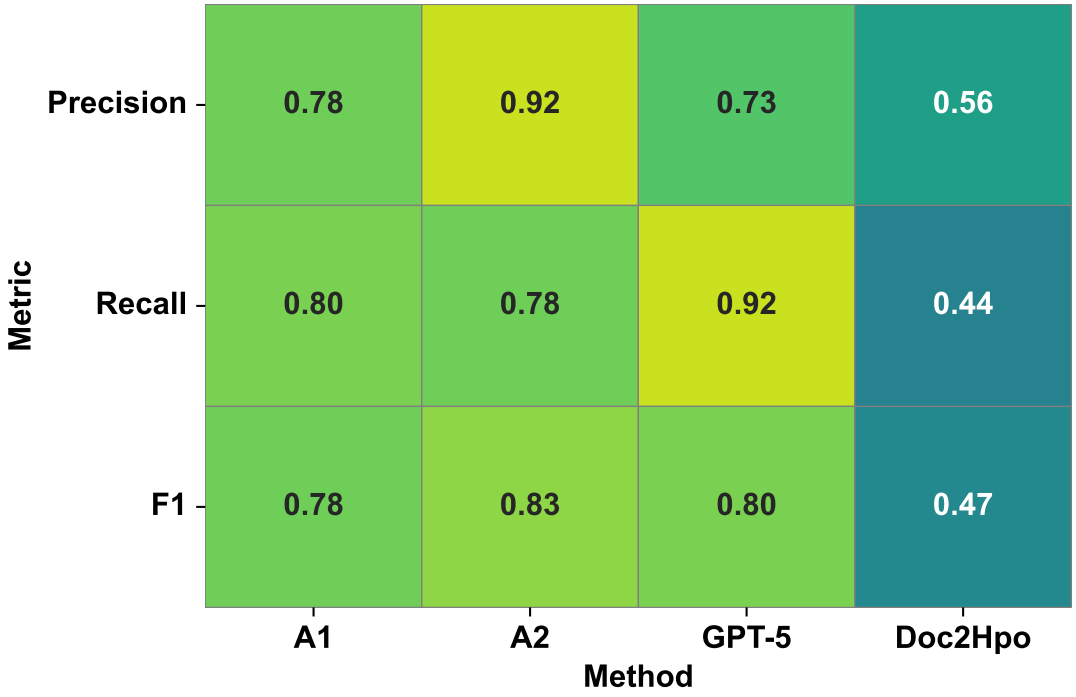

Figure 3. Macro-averaged precision, recall, and F1 scores for the four annotators relative to the gold standard. Scores represent the unweighted mean across 17 phenotypic features, providing a balanced assessment of performance across both common and less frequent phenotypes.

4. Discussion

An important practical motivation for this work is that the manual annotation of neurology notes is time-consuming and inefficient. If large language models could reliably identify and categorize phenotype concepts in free text, large-scale phenotyping of thousands of notes for research and population health applications would become feasible.

We evaluated inter-rater agreement between two human annotators and a large language model (GPT-5) at the note level for phenotype feature identification. Human–human agreement was substantial across most core MS phenotype domains (Table 1). Agreement between the LLM and the human annotators was similarly substantial and extended across most phenotype features (Tables 2 and 3).

Using adjudication of human disagreements, we constructed an adjudicated gold standard annotation set. Relative to this reference, the LLM demonstrated recall that was equal to or higher than human annotators across most phenotype categories (Figure 1). Precision for the LLM was generally comparable to human annotators but was lower in selected domains, including pain, fatigue, bowel, and bladder (Figure 2). These differences reflect a systematically more liberal annotation behavior by the LLM, which tended to infer phenotype presence from indirect evidence (e.g., medication use), whereas human annotators required explicit textual mention. Macro-averaged F1 scores for the LLM were comparable to both human annotators, reflecting a balance between higher recall and modestly reduced precision (Figure 3). In contrast, the dictionary-based Doc2Hpo system lagged behind both human annotators and the LLM, consistent with known limitations of rule-based concept recognition approaches [20,21].

Together, these findings suggest that large language models can approximate clinician-level interpretation of narrative clinical text in a zero-shot setting.

These findings support the emergence of two complementary paradigms for clinical note phenotyping: a fine-grained, span-oriented approach and a coarse-grained, note-oriented approach (Figure 4). Fine-grained methods focus on identifying specific text spans and mapping them to ontology terms and identifiers, enabling detailed linguistic and ontological representation. In contrast, coarse-grained

 doi:10.20944/preprints202604.1173.v1

approaches operate at the level of the complete clinical note, directly inferring the presence or absence of clinically meaningful phenotype domains.

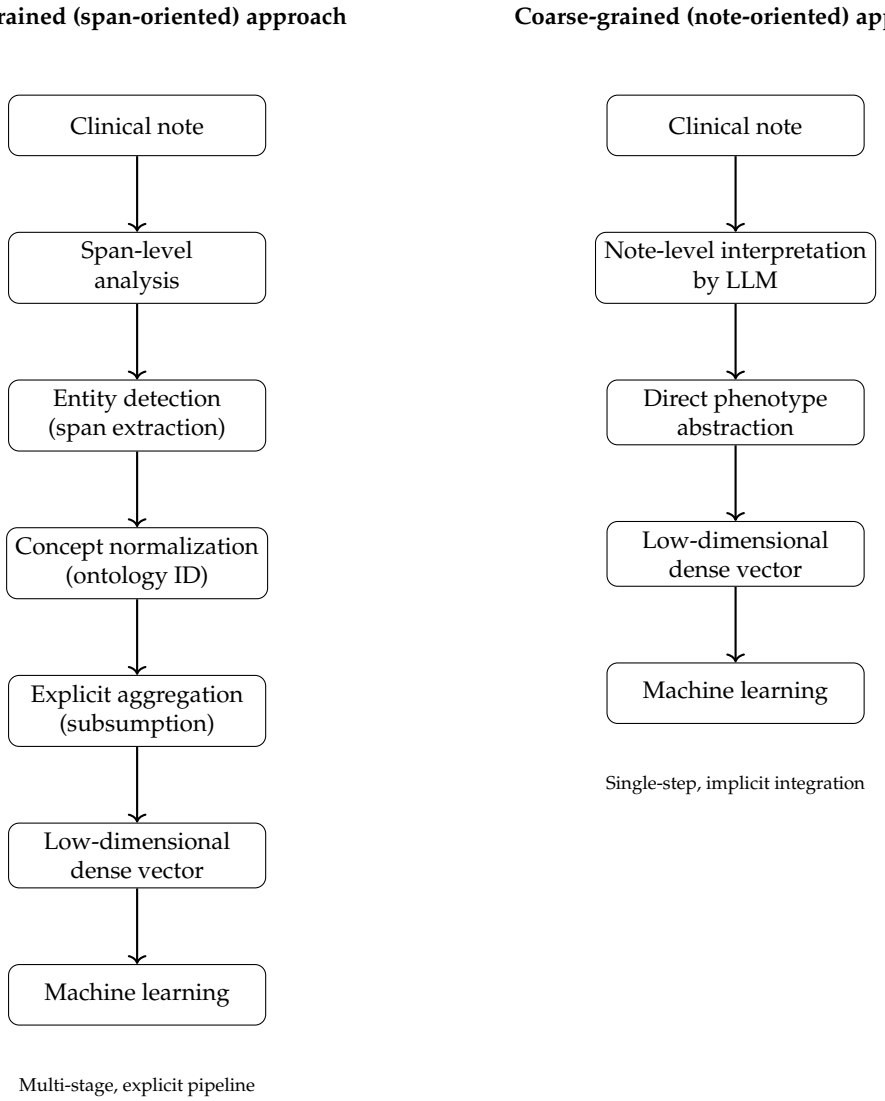

Figure 4. Comparison of two complementary paradigms for generating low-dimensional phenotype representations from clinical text. The fine-grained, span-oriented approach relies on sequential extraction, normalization, and explicit aggregation (subsumption), whereas the coarse-grained, note-oriented approach uses large language models to perform implicit aggregation during direct clinical interpretation. The choice between approaches depends on the intended use case.

These approaches should be viewed as complementary rather than competing, with the choice determined by the intended use case. Fine-grained, span-oriented methods are well suited for applications requiring precise ontology alignment and detailed knowledge representation. Coarse-grained, note-oriented approaches are better suited for large-scale phenotyping and population-level analyses, where the primary objective is robust estimation of phenotype prevalence and patterns across cohorts.

Within this framework, the present study emphasizes the coarse-grained, note-level paradigm as a complementary alternative to traditional fine-grained pipelines. Traditional natural language processing pipelines decompose phenotyping into sequential steps—entity recognition, concept normalization, and aggregation [22–27]. These approaches generate high-dimensional, sparse representations composed of large numbers of fine-granular ontology terms.

However, such representations are often semantically fragmented. Closely related clinical findings—such as dysmetria, hypermetria, and hypometria—are encoded as independent features despite

 doi:10.20944/preprints202604.1173.v1

reflecting a shared underlying cerebellar dysfunction. As a result, shared clinical meaning is not preserved in the feature space, and related findings are treated as separate dimensions rather than manifestations of a unified phenotype. Prior work has addressed this limitation through rule-based subsumption strategies, but these require manually curated mappings and depend on predefined hierarchical relationships [23].

The coarse-grained, note-oriented approach avoids several key limitations inherent to fine-grained pipelines. First, it eliminates the need for precise span boundary determination, which is often ambiguous and inconsistently annotated [28]. Second, it bypasses the requirement for exact mapping of variable and context-dependent clinical language to highly specific ontology terms and identifiers, a process that is both brittle and error-prone [29,30]. Third, it obviates the need for downstream aggregation and subsumption strategies to reduce the dimensionality of highly granular feature spaces, which are often necessary for machine learning applications but introduce additional complexity and sources of variability [23,31].

In contrast to traditional pipelines, large language models appear capable of performing this aggregation implicitly. Rather than extracting individual concepts and subsequently mapping them to higher-level categories, the LLM directly infers clinically meaningful phenotype domains from narrative text. This represents a shift from explicit, rule-based aggregation to implicit, language-driven abstraction, more closely resembling the process by which clinicians synthesize distributed information into diagnostic impressions.

These findings have important implications for clinical natural language processing. By collapsing extraction, normalization, and aggregation into a single interpretive step, LLM-based approaches may improve robustness while reducing implementation complexity and dependence on domain-specific engineering. For downstream applications such as clustering, dimensionality reduction, and machine learning, the ability to directly generate lower-dimensional, clinically meaningful representations may mitigate the challenges associated with sparse, high-dimensional feature spaces [31].

The ability to automate phenotype extraction from narrative MS documentation has important implications for clinical research and health systems analytics. Structured phenotypic data derived from free text could facilitate scalable cohort identification, longitudinal outcome tracking, and real-world evidence generation. For multiple sclerosis, structured capture of features such as gait disturbance, cognitive impairment, and bladder dysfunction may enhance cohort stratification for observational studies and clinical trials.

Although automated systems cannot replace expert clinical judgment, performance approaching human inter-rater agreement indicates that large language models may serve as valuable tools for large-scale phenotyping, particularly when manual chart review is impractical. However, such automated approaches may be unsuitable for certain direct patient care settings that require a high degree of human oversight.

Importantly, these findings suggest that clinical phenotyping may be better framed as a problem of interpretation rather than extraction. Clinically meaningful abstractions can emerge directly from narrative text without first decomposing observations into highly granular ontology terms. For example, a single high-level category such as *hyperreflexia* may be more useful for downstream analysis than multiple fine-grained variants that HPO provides including *generalized hyperreflexia*, *brisk reflexes*, *proximal hyperreflexia*, and *hyperactive deep tendon reflexes* [23].

Limitations

Several limitations should be acknowledged. First, this study included 100 EHR notes from a single institution and time period, which may limit generalizability. Second, annotations were performed by two raters with differing levels of clinical training, which may have influenced agreement patterns. Third, model evaluation was conducted in a zero-shot setting without systematic prompt optimization, and performance may differ under alternative prompting strategies. Fourth, all experiments were conducted using a single large language model (GPT-5), and results may vary with other models or future model versions. Fifth, Doc2Hpo was used as a representative non-LLM comparator; other

 doi:10.20944/preprints202604.1173.v1

automated phenotyping systems may yield different results. Finally, note-level annotation may not capture information available in finer-grained span-level approaches.

Future Directions

Future work should evaluate these findings in larger, multi-institutional datasets and compare large language model approaches directly with traditional supervised NLP pipelines. Additional studies are needed to clarify the relative advantages of fine-grained, span-oriented versus coarse-grained, note-level phenotyping across different use cases. Finally, integration of automated phenotype extraction into clinical data pipelines may enable prospective validation in real-world settings.

4.1. Conclusions

In this study, a large language model achieved performance approaching human inter-rater agreement for note-level MS phenotyping. These findings demonstrate that direct note-level abstraction of clinically meaningful phenotypes from narrative text is feasible and may provide a scalable alternative to traditional extract–map–aggregate pipelines and manual annotation.

More broadly, this work supports the emergence of two complementary paradigms in clinical natural language processing: a fine-grained, span-oriented approach based on explicit extraction and normalization, and a coarse-grained, note-oriented approach based on direct clinical abstraction. Large language models appear to enable the latter by integrating entity recognition, normalization, and aggregation into a single interpretive process.

Within this framework, operations such as roll-up or subsumption are no longer required as explicit post-processing steps but instead emerge implicitly from the model's interpretation. This shift has important implications for how clinical phenotypes are represented in downstream machine learning applications, enabling lower-dimensional, semantically coherent feature representations for large-scale phenotyping and population-level studies [23]. More broadly, this perspective may help guide the selection of automated phenotyping strategies based on the intended use case.

Author Contributions: Conceptualization, P.S. , M.D.C., and D.B.H.; methodology, P.S. and D.B.H. ; formal analysis, P.S. and D.B.H.; writing—original draft preparation, P.S.; writing—review and editing, P.S., D.B.H., and M.D.C.; supervision, D.B.H. and M.D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The use of EHR clinical notes for research was approved by the IRB of the University of Illinois (Protocol 2017-0520Z).

Informed Consent Statement: Patient consent was waived for this retrospective study of existing electronic health record data because the research involved secondary analysis of previously collected clinical information, posed minimal risk to subjects, and used de-identified data under an Institutional Review Board-approved protocol. At hospital admission, patients provide written authorization for the research use of their de-identified health information in IRB-approved studies.

Data Availability Statement: The data underlying this study are not publicly available because they consist of de-identified electronic health record clinical notes that remain subject to institutional and regulatory restrictions on sharing. Data may be made available from the corresponding author on reasonable request and with permission of the University of Illinois, where applicable. The Python code used for data processing and analysis is available from the corresponding author upon reasonable request.

Acknowledgments: This work was performed as part of the UIC Honors College Capstone project by PS for the Honors College of the University of Illinois at Chicago.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202604.1173.v1

Appendix A. Labels

```
PHENOTYPE_LABELS = [
    "Ataxia",
    "Bladder",
    "Bowel",
    "Cognitive",
    "Dysarthria",
    "Dysphagia",
    "Falls",
    "Fatigue",
    "Gait",
    "Hyperreflexia",
    "Pain",
    "Psych",
    "Sensory",
    "Spasticity",
    "Tremor",
    "Vision",
    "Weakness"]
```

Appendix B. Prompt

```
You are labeling a neurology MS note segment for the
PRESENCE/ABSENCE of multiple items.
This is MULTI-LABEL (NOT exclusive):
multiple labels may be present.

LABEL DEFINITIONS:
{LABEL_DEFS}

RULES:
- Use ONLY what is explicitly stated in the TEXT.
- Mark present=true ONLY if there is clear evidence the item applies.
- If present=true: evidence MUST be a short direct quote (<=25 words).
- If present=false: evidence MUST be "" (empty string). detail may be "".
- Do NOT guess. Do NOT use outside knowledge.
- Output MUST include each label at most once.

Return JSON only, matching the schema exactly.
```

Appendix C. Definitions

```
LABEL_DEFS: = {
"Psych": "Psychiatric symptoms (e.g., depression, anxiety, panic, mood disorder)",
"Dysarthria": "Dysarthria, slurred speech, scanning speech.",
"Dysphagia": "Dysphagia, difficulty swallowing, aspiration, feeding tube.",
"Spasticity": ("Spasticity or increased tone/hypertonia/spastic gait.
                Do NOT mark for vague stiffness alone."),
"Hyperreflexia": ("Hyperreflexia/brisk reflexes/increased DTRs (e.g., 3+ or 4+),
                and/or clonus explicitly mentioned."),
"Vision": ( "Visual or eye-movement abnormality (blurred vision, visual field defect,
          diplopia, nystagmus, EOM limitation) explicitly mentioned."),
```

 doi:10.20944/preprints202604.1173.v1


```
"Weakness": "Weakness mentioned (subjective or on exam).",
"Sensory": "Sensory loss/paresthesias/numbness/tingling  by exam or history.",
"Ataxia": "Ataxia/incoordination/dysmetria.",
"Tremor": "Tremor explicitly mentioned.",
"Bladder": "Bladder symptoms (incontinence, urgency, retention, frequency).",
"Bowel": "Bowel symptoms (constipation, diarrhea, incontinence) explicitly mentioned.",
"Cognitive": "Memory/cognitive impairment/brain fog explicitly mentioned.",
"Falls": "Falls explicitly mentioned.",
"Fatigue": "Fatigue explicitly mentioned.",
"Gait": "Gait/balance difficulty; assistive device use (cane/walker/wheelchair).",
"Pain": "Pain explicitly mentioned, including painful cramps.
        burning pain/neuropathic pain."}
```

References

1. Lublin, F.D.; Reingold, S.C.; Cohen, J.A.; Cutter, G.R.; Sørensen, P.S.; Thompson, A.J.; Wolinsky, J.S.; Balcer, L.J.; Banwell, B.; Barkhof, F.; et al. Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology* **2014**, *83*, 278–286.
2. Lublin, F.D.; Reingold, S.C. Defining the clinical course of multiple sclerosis: results of an international survey. *Neurology* **1996**, *46*, 907–911. https://doi.org/10.1212/WNL.46.4.907.
3. Miller, A.E.; Coyle, P.K. Clinical features of multiple sclerosis. *CONTINUUM: Lifelong Learning in Neurology* **2004**, *10*, 38–73.
4. Ford, H. Clinical presentation and diagnosis of multiple sclerosis. *Clinical Medicine* **2020**, *20*, 380–383.
5. Kurtzke, J.F. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* **1983**, *33*, 1444–1452.
6. Davis, M.F.; Sriram, S.; Bush, W.S.; Denny, J.C.; Haines, J.L. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *Journal of the American Medical Informatics Association* **2013**, *20*, e334–e340.
7. Rae-Grant, A.; Day, G.S.; Marrie, R.A.; Rabinstein, A.; Cree, B.A.; Gronseth, G.S.; Haboubi, M.; Halper, J.; Hosey, J.P.; Jones, D.E.; et al. Practice guideline recommendations summary: disease-modifying therapies for adults with multiple sclerosis: report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology* **2018**, *90*, 777–788.
8. Correia, I.; Bernardes, C.; Cunha, C.; Nunes, C.; Macário, C.; Sousa, L.; Batista, S. Picturing the Multiple Sclerosis patient journey: a symptomatic overview. *Journal of Clinical Medicine* **2024**, *13*, 5687.
9. Swetlik, C.; Bove, R.; McGinley, M. Clinical and research applications of the electronic medical record in multiple sclerosis: a narrative review of current uses and future applications. *International Journal of MS Care* **2022**, *24*, 287.
10. Qi, J.; Luo, L.; Yang, Z.; Wang, J.; Zhou, H.; Lin, H. An improved method for phenotype concept recognition using rich HPO information. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2024, pp. 1135–1140.
11. Dahdul, W.; Manda, P.; Cui, H.; Balhoff, J.P.; Dececchi, T.A.; Ibrahim, N.; Lapp, H.; Vision, T.; Mabee, P.M. Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. *Database* **2018**, *2018*, bay110.
12. Oommen, C.; Howlett-Prieto, Q.; Carrithers, M.D.; Hier, D.B. Inter-rater agreement for the annotation of neurologic signs and symptoms in electronic health records. *Frontiers in Digital Health* **2023**, *5*, 1075771.
13. Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* **1960**, *20*, 37–46.
14. McHugh, M.L. Interrater reliability: the kappa statistic. *Biochemia medica* **2012**, *22*, 276–282.
15. OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* **2023**.
16. Yang, X.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Flores, M.G.; Zhang, Y.; Magoc, T.; et al. Gatortron: A large language model for clinical natural language processing. *MedRxiv* **2022**, pp. 2022–02.
17. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180.

 doi:10.20944/preprints202604.1173.v1

18. Groza, T.; Köhler, S.; Doelken, S.; Collier, N.; Oellrich, A.; Smedley, D.; Couto, F.M.; Baynam, G.; Zankl, A.; Robinson, P.N. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database* **2015**, *2015*, bav005.
19. Munzir, S.I.; Hier, D.B.; Oommen, C.; Carrithers, M.D. A large language model outperforms other computational approaches to the high-throughput phenotyping of physician notes. In Proceedings of the AMIA Annual Symposium Proceedings, 2025, Vol. 2024, p. 838.
20. Liu, C.; Kury, F.; Li, Z.; Ta, C.N.; Wang, K.; Weng, C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic Acids Research* **2019**, *47*, W566 – W570. https://doi.org/10.1093/nar/gkz386.
21. Yang, J.; Liu, C.; Deng, W.; Wu, D.; Weng, C.; Zhou, Y.; Wang, K. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns* **2023**, *5*. https://doi.org/10.1016/j.patter.2023.100887.
22. Azizi, S.; Hier, D.B.; Wunsch II, D.C. Enhanced neurologic concept recognition using a named entity recognition model based on transformers. *Frontiers in Digital Health* **2022**, *4*, 1065581.
23. Hier, D.B.; Yelugam, R.; Carrithers, M.D.; Wunsch III, D.C. The visualization of Orphadata neurology phenotypes. *Frontiers in Digital Health* **2023**, *5*, 1064936.
24. Berkowitz, J.S.; Srinivasan, A.; Cortina, J.M.A.; Fatapour, Y.; Tatonetti, N.P. Biomedical text normalization through generative modeling. *Journal of Biomedical Informatics* **2025**, *167*, 104850.
25. Azam, I.; Jiang, K.; Bernard, G. Normalizing Health Concepts with Biomedical Embedding and LLMs. In Proceedings of the Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), 2026, pp. 180–190.
26. Shlyk, D.; Montanelli, S.; Mesiti, M.; Hunter, L. Mind Your Steps in Biomedical Named Entity Recognition: First Extract, Tag Afterwards. In Proceedings of the Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), 2026, pp. 127–141.
27. Fu, S.; Chen, D.; He, H.; Liu, S.; Moon, S.; Peterson, K.J.; Shen, F.; Wang, L.; Wang, Y.; Wen, A.; et al. Clinical concept extraction: a methodology review. *Journal of biomedical informatics* **2020**, *109*, 103526.
28. Baddour, M.; Paquelet, S.; Rollier, P.; Tayrac, M.; Dameron, O.; Labbé, T. Phenotypes Extraction from Text: Analysis and Perspective in the LLM Era. In Proceedings of the 2024 IEEE 12th International Conference on Intelligent Systems (IS), 2024, pp. 1–8. https://doi.org/10.1109/is61756.2024.10705235.
29. Garcia, B.T.; Westerfield, L.; Yelemali, P.; Gogate, N.; Rivera-Munoz, E.A.; Du, H.; Dawood, M.; Jolly, A.; Lupski, J.R.; Posey, J. Improving automated deep phenotyping through large language models using retrieval-augmented generation. *Genome Medicine* **2025**, *17*. https://doi.org/10.1186/s13073-025-01521-w.
30. Hier, D.B.; Do, T.S.; Obafemi-Ajayi, T. A simplified retriever to improve accuracy of phenotype normalizations by large language models. *Frontiers in Digital Health* **2025**, *7*, 1495040.
31. Wunsch III, D.C.; Hier, D.B. Subsumption reduces dataset dimensionality without decreasing performance of a machine learning classifier. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021, pp. 1618–1621.