Leveraging Large Language Models to Derive Multiple Sclerosis Progression Assessments from Clinical Notes: A Feasibility Study

Sy Hwang[1†], Sunil Thomas[1], Heather Williams[1], Tom Hutchinson[1], Emily Schriver[1], Ashley Batugo[1], Amit Bar-Or[2], Vishakha Sharma[4], Frederik Buijs[3], Christopher Perrone[2], and Danielle Mowery[1]

[1]*Institute for Biomedical Informatics, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA USA*

[2]*Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA USA*

[3]*F. Hoffmann-La Roche, Basel, Switzerland*
[4]*Roche Diagnostics, Santa Clara, CA USA*

[†]*Corresponding author e-mail: sy.hwang@pennmedicine.upenn.edu*

Ascertainment of multiple sclerosis (MS) progression is important for informing clinical care decisions and supporting biomedical research. However, the details to infer a patient's MS progression status are locked within clinical notes. In this feasibility study, we assessed the feasibility of developing and validating a large language model (LLM)-based EDSS and FS classifier for ascertaining MS progression from clinical notes.

Keywords: Multiple Sclerosis; Large Language Models; Natural Language Processing.

1. Introduction

Multiple sclerosis (MS), a chronic inflammatory disorder of the brain and spinal cord, is estimated to affect 2.9 million people worldwide and 1 million people in the United States.[1] The condition is characterized by acute episodes of symptoms from which a patient often recovers well, referred to as relapses, and gradual accrual of disability over time, referred to as progression of disability. The presence of these features has formed the basis for the commonly used clinical course descriptors which refer to an individual's clinical phenotype. For instance, patients typically experience relapses early in their clinical referred to as relapsing-remitting multiple sclerosis (RRMS), many of them later manifest with gradual progression of disability that has been referred to as secondary progressive multiple sclerosis (SPMS). Others do not have evident relapses and develop gradual progression of disability from clinical onset, a phenotype referred to as primary progressive multiple sclerosis (PPMS).

While relapses are relatively easy to identify along a patient's disease course due to typically more rapid change in symptoms, MS progression can be more difficult to discern because accrual of disability develops more slowly in subtle ways, often taking many months

© 2025 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

or years to appreciate. Yet, it is critical to better identify progression as treatment approaches are evolving and can be different. The Expanded Disability Status Scale (EDSS), developed by John Kurtzke in 1983, is a validated clinical assessment tool for measuring disability and its progression in MS.[2] It assigns scores for the degree of neurologic impairment across eight domains called functional systems (FS). These systems include functions in the `visual`, `brainstem`, `pyramidal (strength)`, `cerebellar (coordination)`, `sensory`, `bowel and bladder`, `cerebral (cognitive)`, and `ambulatory (walking)` domains. Each FS contributes to a calculation of overall EDSS score, which ranges from 0 (normal) to 10 (death) with intervals of 0.5. The EDSS is used as a standard outcome measure for clinical trials, and a change of +1.0 for EDSS scores between 0 and 5.5 or a change of +0.5 for EDSS scores between 6 and 10 is deemed evidence of 'clinically significant' progression. The time needed to determine each FS score and the overall EDSS score can pose a challenge for measurement in routine care. There are heuristics that can be used to more easily derive certain EDSS scores (such as the use of a cane resulting in an EDSS of 6.0), but this approach is limited.

Scores can also be retrospectively derived based on the narrative and description of the neurological exam in clinic notes. Natural language processing (NLP) may aid automatic computation of FS and EDSS scores to support both routine clinical care as well as clinical research studies of treatment efficacy. Several studies have explored automated recovery of EDSS from routinely collected data with NLP. Yang et al. compared rule-based, convolutional neural network (CNN), and a combined rule-based+CNN model and found the combined approach performed best for total EDSS (accuracy = 0.90; F1 = 0.83), while FS subscore performance was lower, especially when subscores were not explicitly documented in notes.[3] Alves et al. validated an XGBoost-based estimator on neurologist notes, reporting PPV 0.85 and NPV 0.85[4] . D'Costa et al. introduced MS-BERT and an MS severity classifier, showing transformer-based embeddings substantially improved EDSS macro-F1 of 0.88 and offered gains across subscores vs. Word2Vec and rule-based baselines, yet FS extraction remained the harder task.[5] Beyond note-level NLP, Davis et al. demonstrated early feasibility of rule-driven EDSS extraction from EHR text at scale, motivating subsequent learning-based approaches.[6] Muros-Le Rouzic et al. derived a claims-based proxy for EDSS and validated it against chart EDSS.[7] While granular classification was weak, broader severity groupings, particularly EDSS $\geq$ 6.0, showed much better agreement, underscoring that text-rich notes remain the more precise substrate for fine-grained EDSS. We hypothesize that large language models might improve upon performance for FS in clinical notes because of their ability to reason over heterogeneous phrasing and cross-sentence context, integrating implicit cues with section-aware constraints to map narrative evidence to FS constructs even when explicit subscores are absent.

In the long-term, our objective is to develop a robust and highly accurate, multimodal AI algorithm for determining MS progression of disability over time. For this short-term pilot study, our objective is two-fold: (1) determine the feasibility of automatically deriving MS disability progression from clinical notes leveraging large language models (LLMs) and (2) evaluate how precisely EDSS progression scores can be derived from clinical notes.

2. Methods

This retrospective, feasibility study was approved by the University of Pennsylvania Institutional Review Board.

2.1. *Cohort*

We identified patients with MS from the Penn Neuroimmunology Registry with a diagnosis of MS by a neurologist according to McDonald criteria (revised) treated between calendar years 2024-2025 (n=239 patients with 288 progress notes). The mean number of notes per unique patient encounter was 1.2 ± 0.41.

2.2. *Multiple Sclerosis Progression*

We developed LLM-based classifiers for extracting and encoding two MS-related progression indicators: the standardized and quantified neurological examination and assessment of Kurtzke's `Functional Systems (FS)` and `Expanded Disability Status Scale (EDSS)` from the clinical notes. The complete definitions and logic for computing the EDSS score can be found in Kappos et al.[8] Below, we offer the following more detailed, descriptive summary for FS and EDSS scoring:

(1) **Functional Status (FS)** include subscores based on 8 neurological functions that determine a patient's ability to perform daily activities - basic and complex - necessary for independent living.[9]

   (a) `Visual Functions` include assessments of visual acuity, visual fields, scotoma, and disc pallor with functional system scores ranging from 0=normal to 6=grade 5 plus maximal visual acuity of better eye of 20/60 (0.33) or less,
   (b) `Brainstem Functions` include assessments of extraocular movement impairment, nystagmus, trigeminal damage, facial weakness, hearing loss, dysarthria, dysphagia, and other cranial nerve functions with function system scores ranging from 0=normal to 5=inability to swallow or speak,
   (c) `Pyramidal Functions` include assessments of reflexes, limb strength, functional tests, limb spasticity, gait spasticity, overall motor performance with function system scores ranging from 0=normal to 6=tetraplegia defined as British Medical Research Council (BMRC) grade 0 or 1 in all muscle groups of the upper and lower limbs,
   (d) `Cerebellar Functions` include assessments of head tremor, truncal ataxia, limb ataxia, tandem walking, gait ataxia, Romberg test, other cerebellar tests with function system scores ranging from 0=normal to X=pyramidal weakness BMRC grade 3 or worse in limb strength) or sensory deficits interfere with cerebellar testing,
   (e) `Sensory Functions` include assessments of superficial sensation, vibration sense, position sense, Lhermitte's sign, and paraesthesiae with function system scores ranging from 0=normal to 6=sensation essentially lost below the head,
   (f) `Bowel and Bladder Functions` include assessments of urinary hesitancy and retention, urinary urgency and incontinence, bladder catheterization, bowel dysfunction,

and sexual dysfunction with function system scores ranging from 0=normal to 6=loss of bowel and bladder function,

(g) `Cerebral Functions` include assessments of depression and euphoria, decrease in mentation, and fatigue with function system scores ranging from 0=normal to 5=dementia,

(h) `Ambulation Functions` include assessments of distance and time reported by patient and use of devices with function system scores ranging from 0=unrestricted to 12=essentially restricted to bed or chair or perambulated in wheelchair, but out of bed most of day; retains many self-care functions; generally has effective use of arms (EDSS 8.0).

(2) **Expanded Disability Status Scale (EDSS)** is a standardized assessment tool for determining the degree of disability in patients with MS that is calculated based on the individual FS subscores. The EDSS scores range from 0=normal neurological exam (all FS grade 0) to 10=death due to MS.[9] Because the ambulatory is the largest scaled FS score, it tends to have drive the overall EDSS score. Therefore, the ambulation score can be directly mapped to the EDSS score in a few limited cases illustrated below.

- 0=unrestricted
- 1 = fully ambulatory
- 2 = Walks $\geq$ 300 m, but $<$ 500 m unassisted; mild gait impairment (EDSS 4.5 or 5.0)
- 3 = Walks $\geq$ 200 m, but $<$ 300 m unassisted; noticeable limitation (EDSS 5.0)
- 4 = Walks $\geq$ 100 m, but $<$ 200 m unassisted; may require minimal aid (EDSS 5.5)
- 5 = walking range $< 100m$ w/o assist (EDSS 6.0)
- 6= unilateral assist, $\geq$ 50 m (EDSS 6.0)
- 7= bilateral assist, $\geq$ 120 m (EDSS 6.0)
- 8= unilateral assist, $<$ 50 m (EDSS 6.5)
- 9= bilateral assist, $\geq$ 5 m, but $<$ 120 m (EDSS 6.5)
- 10 = wheelchair w/o help (EDSS 7.0)
- 11 = wheelchair w help (EDSS 7.5)
- 12 = bed or chair bound (EDSS 8.0)

2.3. *Prompts*

We used OpenAI's GPT-4.1 model, which we have historically demonstrated its ability to complete information extraction and patient phenotyping tasks using clinical notes.[10] To reduce the model's stochasticity and improve reproducibility, we set parameters of temperature at 0.1 and top_p at 1. For this feasibility study, we opted to use zero-shot learning after several rounds of curating and making iterative changes to the prompts. We created `system_prompts` and `task_prompts` for executing FS and EDSS processing and classification steps.

```
system_prompt = As a neurologist at a health system, you have been tasked
with reviewing clinical notes for evidence of progression of multiple
sclerosis, an autoimmune disease of the brain and spinal cord. You will be
scoring functional system subscores for patients. For each measure,
analyze the narrative and determine if the type is explicitly and
undeniably present or absent. The clinical note you are assessing can be
lengthy, so focus only on the relevant parts for the task
```

2.3.1. *FS prompt and processing task*

For each functional system, the LLM was instructed carefully review the clinical note and provided more detailed logic to assign a score based on information documented within the narrative. Below, we provide truncated examples for each FS subscore.

FS_task_prompt = For each functional system, carefully review the clinical note and assign a score based on the following criteria:

(1) VISUAL SUBSCORE:

- Score 0: if none of the below are true ...
- Score 5: if there is mention of no light perception (NLP) ...

(2) BRAINSTEM SUBSCORE:

- Score 0: Cranial nerves described as normal or no abnormalities mentioned ...
- Score 5: Unable to speak or swallow, requires enteral nutrition or PEG tube ...

(3) PYRAMIDAL SUBSCORE:

- Score 0: if none of the below are true ...
- Score 6: if there is mention of tetraplegia, all limbs with a grade of 0 or 1 ...

(4) CEREBELLAR SUBSCORE:

- Score 0: if the cerebellar exam is normal or none of the below are true ...
- Score 4: if severe truncal AND limb ataxia is present ...

(5) SENSORY SUBSCORE:

- Score 0: if the sensory exam is described as normal or none of the below are true ...
- Score 6: 6 if complete loss of all sensation in 3-4 limbs (neck down) ...

(6) BOWEL/BLADDER SUBSCORE:

- Score 0: if none of the below are true ...
- Score 6: if complete loss of bladder and bowel control ...

(7) AMBULATION SUBSCORE:

- Score 0: if gait exam is normal or none of the below are true ...
- Score 12: if uses a motorized chair ...

(8) CEREBRAL SUBSCORE:

- Score 0: if none of the below are true, exclude mood related symptoms and also score 0 if there is improvement in a cognitive symptom ...
- Score 5: if there is mention of dementia ...

2.3.2. *EDSS prompt and processing*

For the EDSS task, the LLM was instructed to ascertain the EDSS score from the FS scores using the `EDSS_task_prompt` and provided more detailed post-processing steps that incorporated instructions from Kappos et al.[8]

```
EDSS_task_prompt = You are an expert neurologist specializing in multiple
sclerosis (MS), a chronic autoimmune disease affecting the central nervous
system. One of your key responsibilities is accurately determining a
patient's level of disability using the Expanded Disability Status Scale
(EDSS). The EDSS is the gold standard measure of disability in MS, ranging
from 0 (normal neurological examination) to 10 (death due to MS). This
scale integrates impairments across multiple functional systems into a
single score that reflects overall disability. Your expertise is required
to translate a patient's functional system subscores into an accurate EDSS
rating. These subscores come from a comprehensive neurological examination
and measure impairment in eight key domains: ambulation (walking ability),
vision, brainstem function, pyramidal function (motor strength),
cerebellar function (coordination), sensory function, bowel/bladder control,
and cerebral/mental function. Your goal is to apply the established EDSS
calculation rules consistently and accurately.
```

Using the output from the eight predicted functional system (FS) subscores, an LLM was instructed to deterministically derive an EDSS score (0.0–9.0). Two additional rules must be applied to correctly ascertain clinically-valid EDSS scores. For visual and bowel/bladder FS, scores are adjusted so that these symptoms do not disproportionately contribute to the EDSS score. Use of an assistive device suggests a significant degree of disability and therefore some ambulatory FS scores determine a specific EDSS score regardless of other FS values. The LLM outputs were used to compute the EDSS score using the following steps in order:

(1) Pre-convert `visual` and `bowel/bladder` FS scores to a lower scale
(2) Convert `ambulatory` FS scores directly to EDSS values when possible.

The model was prompted to (i) articulate a concise chain of thought for each of the eight functional-system domains, (ii) assign the corresponding ordinal score, and (iii) return the predictions in a simple, consistently formatted JSON object. Requiring the model to "reason first, a nswer s econd" e nabled i nspection o f i ts i ntermediate r ationale a nd h as b een shown to improve zero-shot classification p erformance i n c linical l anguage t asks.[11,12] H andling one domain at a time limited cross-domain interference and keeps the prompt length manageable for long progress notes. A rigid output schema ensured that the predictions are machine-readable without post-processing, which simplified i ntegration w ith t he e valuation pipeline and downstream analytic workflows.

2.4. *FS Classifier Performance*

We assessed the performance of the FS classifier u sing a ccuracy a s w ell a s p recision, recall, and both weighted and macro F1.

2.5. *EDSS Classifier Performance*

We understand that there is imprecision in classifying EDSS scores from FS scores due to lack of specificity in the measurement and incorrect inference among other reasons. However, not all misclassifications are alike. For example, the documentation of the `ambulatory` FS score may not include distance walked to precisely classify lower scores. Therefore, we devised several lenient match criteria for capturing and assessing performance based on ordinal closeness (adjusts for small numeric differences in documented precision such as off by n) and semantic relatedness (adjusts for similarities in clinical definitions such as use of assistive device).

(1) Exact match: both expert and LLM values match exactly
(2) Lenient match by 1: the LLM value is within 1 point of the expert value
(3) Lenient match by 2: the LLM value is within 2 points of the expert value
(4) Lenient match by category: the LLM value is an exact or falls within a range of values associated with need for assistance. These ranges were as follows: *no assist* ranges from $0 \leq 5.5$, *unilateral assist* is 6, *bilateral assist* is 6.5, and *wheelchair* ranges from $7 \leq 8.5$.

We computed the following agreement and error metrics:

(1) Quadratic Kappa (κ): quantifies agreement by penalizing errors based on order and distance between the true and predicted scores; penalizes larger disagreements more heavily.
(2) Mean Average Error (MAE): measures the average absolute distance between true and predicted scores; interprets error rate by treating all errors linearly.
(3) Root Mean Square Error (RMSE): measures the square root of the mean squared differences between the true and predicted scores; penalizes larger errors more heavily than MAE, making it useful for class imbalance and comparing costly errors and outliers.

3. Results

We developed LLM-based classifiers for extracting and inferring FS and EDSS scores from clinical notes. We report classification performance.

3.1. *FS Classifier Performance*

In **Table 1**, we observed high overall accuracy and weighted recall across FS subtypes ranging from 0.787 (`cerebral`) to 0.971 (`visual`). Weighted F1-scores ranged from moderate at 0.784 (`ambulation`) to high at 0.970 (`visual`); in contrast, macro F1-scores ranged from low at 0.295 (`ambulation`) to moderate at 0.724 (`visual`). The majority class for each FS subscore is 0.

In **Figure 1**, we show the distribution of FS subscores by the generated by the expert and LLM. We observe similar trends within subscore values, e.g., both predict similar prevalence across subscores and the LLM under predicts the majority class of 0 and overpredicts value 1.

In **Figure 2**, we assessed the difference between true, expert-generated FS score values and predicted, LLM-generated FS score values. We observe concordant median score values of 0 for all functional system scores and most exactly match across FS types. Most of the predicted LLM values are off by no more than 2 points from the true value.

Table 1: Performance metrics for FS subtype classification

FS subtype	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
			Weighted			Macro	
`Ambulation`	0.799	0.776	0.799	0.784	0.275	0.362	0.295
`Bowel/Bladder`	0.782	0.876	0.782	0.803	0.476	0.447	0.413
`Brainstem`	0.900	0.930	0.900	0.911	0.495	0.561	0.511
`Cerebellar`	0.879	0.892	0.879	0.878	0.691	0.619	0.631
`Cerebral`	0.787	0.849	0.787	0.801	0.640	0.664	0.641
`Visual`	0.971	0.970	0.971	0.970	0.708	0.744	0.724
`Pyramidal`	0.812	0.815	0.812	0.806	0.456	0.436	0.440
`Sensory`	0.895	0.901	0.895	0.892	0.660	0.638	0.642

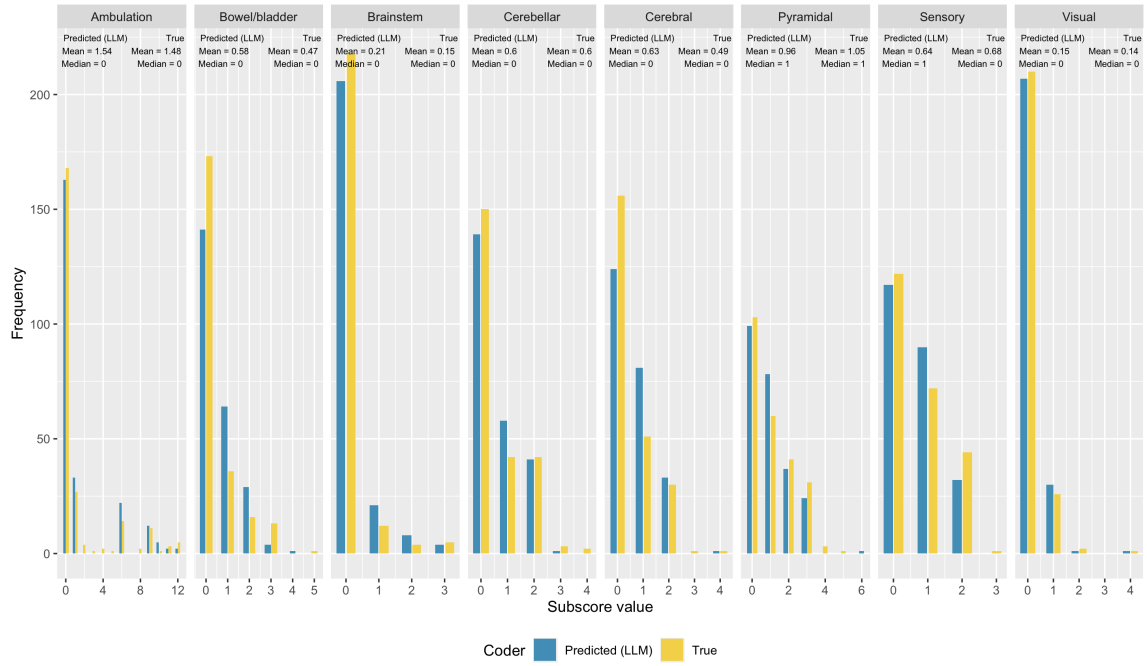

Fig. 1: Distribution of true and predicted FS scores

3.2. *EDSS Classifier Performance*

In **Figure 3**, we performed a sensitivity analysis using exact and lenient match criteria for computing expert-LLM agreement. For exact IAA, EDSS performance ranged from low (score 2.5 at 0.33) to high (score 7 at 1.0). 1 of 16 EDSS subscores achieved 100% agreement. For lenient by 1 IAA, 10 of 16 (62.5%) EDSS subscores achieved 100% agreement; for lenient by 2 IAA, 12 of 16 (75%) EDSS subscores achieved 100% agreement. Match by category produced a high overall of 0.974 and moderate to high IAA: no assist of 0.980, unilateral assist of 0.667, bilateral assist of 0.846, and wheelchair of 1.00.

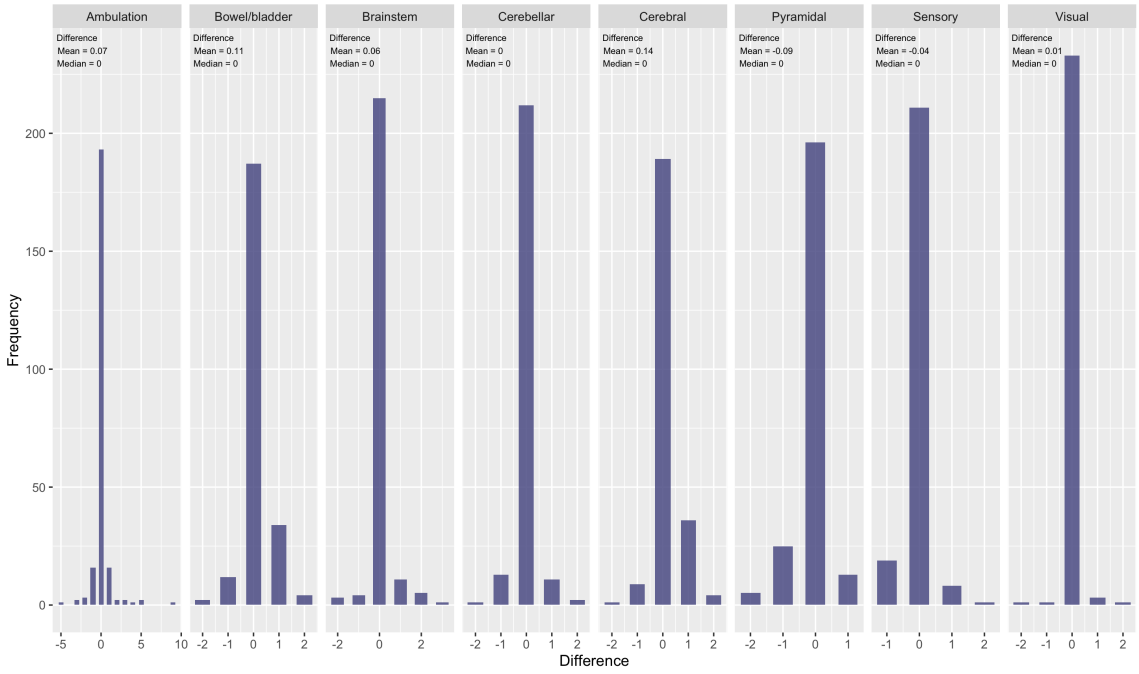

Fig. 2: Differences between true and predicted FS scores

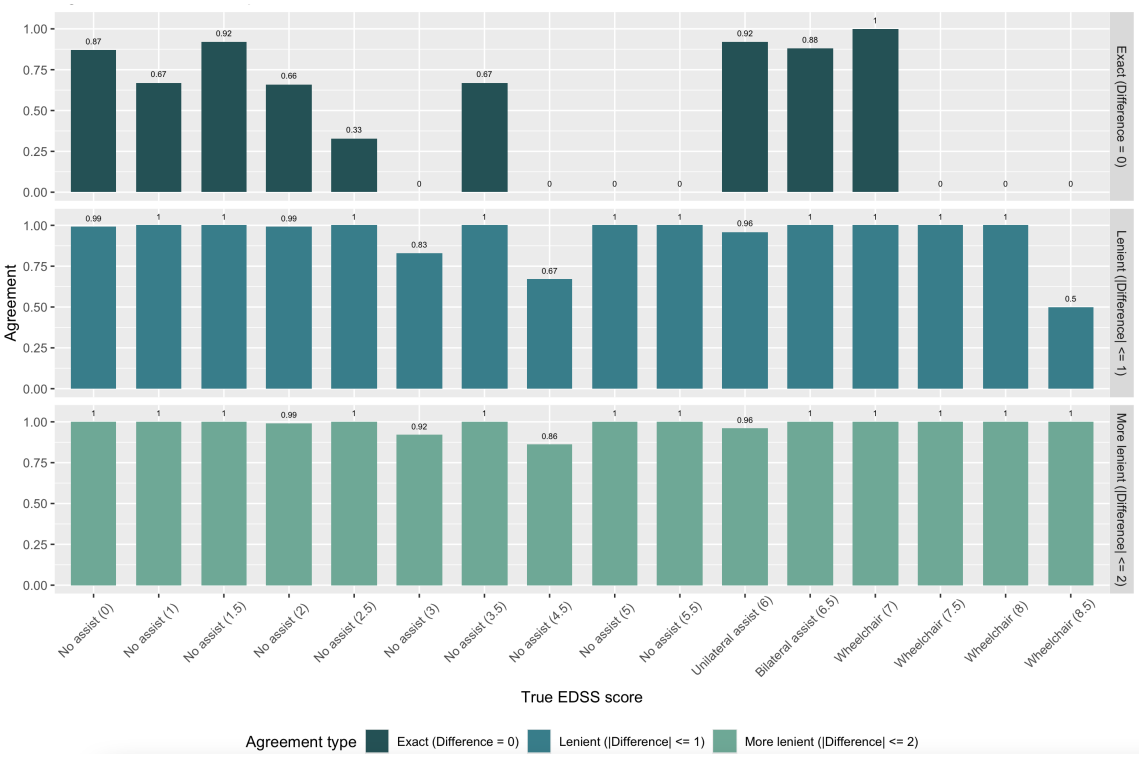

Fig. 3: Agreement between true and predicted EDSS scores varied by match criteria

In **Figures 4 and 5 - Appendix**, we report the confusion matrices with κ, MAE, and RMSE for FS and EDSS scores. κ ranged from moderate (`brainstem` at 0.621) to high

(ambulation at 0.942) for subscores and was high for EDSS at 0.942. MAE ranged from 0.13 for bowel/bladder to 0.34 for ambulation subscores and was 0.34 for EDSS. RMSE ranged from 0.39 for bowel/bladder to 1.04 for ambulation for subscores and was 0.70 for EDSS.

4. Discussion

We conducted a feasibility study leveraging LLMs to encode FS and EDSS scores from clinical notes and an expert-LLM sensitivity analysis to understand imprecise classification of EDSS.

4.1. *FS Classifier Performance*

The language model correctly assigned FS scores in most domains. Five of eight domains exceeded exact accuracy of 0.90. The lower macro-average numbers reflected the class imbalance and the difficulty of long-tailed labels. Visual, sensory, and cerebellar domains achieved macro-F1 above 0.70, whereas ambulation and pyramidal lagged below 0.40. Performance differences follow clinical documentation patterns. Visual loss and cerebellar disorders are usually described with distinct terms such as "optic neuritis" which the model captures reliably. Ambulation scoring depends on walking distance and aid usage, information often omitted or expressed qualitatively. Pyramidal signs occupy a similar gray zone; progress notes may list upper motor neuron findings without quantifying weakness, making precise ordinal mapping challenging.

4.2. *EDSS Classifier Performance*

We observed that most EDSS misclassifications could be captured with more lenient match criteria. Error analysis shows that most residual errors arise at the scale's extremes. In the very mild range (EDSS 0.5–2.5) the model tends to underestimate severity by one half-point, likely because phrases such as "walks independently but tires easily" are interpreted as normal ambulation, demonstrating that underspecified and small variations in ambulation phrasing can lead to whole-point shifts. In the wheelchair-bound range (EDSS $\geq$ 7.0) we observe a few adjacent misclassifications between 7.0, 7.5, and 8.0, but all predictions remain within the acceptable tolerance band.

Because EDSS is an ordinal measure, the clinical impact of a prediction error depends on its magnitude. Misreading a note scored by clinicians as EDSS 6.5, where the patient needs bilateral support to walk, as 2.0, which is indicated only minimal disability, could delay escalation of therapy or understate safety concerns. By contrast, confusing 4.0 with 5.0 seldom changes management. Cognitive error, rather than lack of knowledge, underlies many medical mistakes, and evaluation metrics that ignore the distance between ordinal classes can hide systems that make infrequent yet clinically significant large-magnitude errors. This concern is particularly relevant in the context of bridging AI with human expertise, where misjudgment of disability trends directly distorts decisions about disease progression and treatment response.

4.3. *Related Works*

In comparison to Yang et al.'s highest performing classifiers for FS subscores, our LLM-based solution produced higher F1-scores for `cerebellar`, `cerebral`, `visual`, and `sensory`. However, the LLM-based EDSS classifier performance did not outperform the rules/CNN-based classifier. We hypothesize that poor EDSS classification is partially driven by FS classification as well as insufficient ambulatory score details, e.g., how far a patient can walk in feet documented in the clinical notes.

4.4. *Limitations and Future Work*

This work presents a feasibility evaluation situated alongside prior efforts (Yang et al.), with the primary objective of establishing a clear empirical reference point for automated EDSS and FS extraction from routine notes. Several limitations merit emphasis. Our cohort is modest and drawn from a single institution, which raises the risk of overfitting, particularly given iterative prompt refinement, and limits generalizability. We will mitigate this by expanding the dataset across additional neurologists, encounter types, and less prevalent MS phenotypes; freezing explicit train/dev/test splits with documented seeds and access controls; and employing automated prompt optimization that searches over instruction variants and exemplars to reduce ad-hoc tuning and quantify sensitivity to prompt choices. Labeling in the present study relied in part on a single annotator, which constrains ceiling performance and injects unavoidable uncertainty; the next phase includes multi-rater annotation with adjudication on a stratified subset and reporting of inter-rater agreement.

Methodologically, our zero-shot framing focuses on clinical utility and transparency but remains incremental. Two directions are in progress to strengthen novelty and fit to the task. First, we are conducting targeted error analyses and ablation studies to isolate failure modes and their impact on EDSS and FS outputs. Second, we are exploring a presence-only or maximum-entropy formulation in which the LLM serves as a high recall evidence extractor and a constrained probabilistic layer estimates FS abnormalities and ambulation anchors without treating "no mention" as negative. An ordinal mapping with clinical consistency constraints could then yield EDSS scores. This hybrid approach with LLMs has the potential to improve calibration, reduce silent false negatives, and better respect the ordinal structure of disability scores.

Finally, because EDSS is computed from multiple functional subsystems, errors in any single domain can propagate to the composite score. To curb this, future iterations will incorporate complementary modalities to cross-validate narrative inferences, designate exact-match EDSS as the primary endpoint with tolerance-band results reported as secondary analyses, and extend analyses beyond point estimates to longitudinal trajectories, enabling detection of clinically meaningful changes over time. We emphasize that tolerance bands are not a substitute for exact agreement. They are reported to reflect the known resolution of EDSS assignment in routine practice and to distinguish near-misses from clinically divergent errors. In our error analysis, most disagreements were adjacent and preserved the underlying category, suggesting limited clinical impact. With a larger, more diverse corpus, multi-site validation, multi-rater labels, and principled prompt/model tuning, we aim to deliver a more robust, generalizable

system suitable for prospective evaluation while preserving the practicality and transparency that motivated this feasibility stage.

5. Conclusion

Automating EDSS and FS extraction can lessen clinician documentation burden, unlock historical data for large-scale outcome studies, and provide up-to-date disability estimates to decision support tools. This pilot investigation confirms the technical feasibility of deriving detailed MS disability progression measures from routine neurology notes with a single prompt-based large language model. The model achieved strong FS performance, with accuracy spanning $0.782 – 0.971$ and weighted F1 ranging from 0.784 to 0.970 across the eight neurologic domains. Although overall EDSS exact match agreement was lower, applying clinically-meaningful tolerance bands raised categorical agreement to 0.974, indicating that most disagreements occur within ranges unlikely to alter therapeutic decision making.

6. Acknowledgments

This project was partially funded by a grant from F. Hoffmann-La Roche, Basel, Switzerland.

References

1. L. M. Nelson, M. T. Wallin, R. A. Marrie, W. Culpepper, A. Langer-Gould, J. Campbell, S. Buka, H. Tremlett, G. Cutter, W. Kaye *et al.*, A new way to estimate neurologic disease prevalence in the united states: illustrated with ms, *Neurology* **92**, 469 (2019).
2. J. F. Kurtzke, Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss), *Neurology* **33**, 1444 (1983).
3. Z. Yang, C. Pou-Prom, A. Jones, M. Banning, D. Dai, M. Mamdani, J. Oh, T. Antoniou *et al.*, Assessment of natural language processing methods for ascertaining the expanded disability status scale score from the electronic health records of patients with multiple sclerosis: algorithm development and validation study, *JMIR Medical Informatics* **10**, p. e25157 (2022).
4. P. Alves, E. Green, M. Leavy, H. Friedler, G. Curhan, C. Marci and C. Boussios, Validation of a machine learning approach to estimate expanded disability status scale scores for multiple sclerosis, *Multiple sclerosis journal - experimental, translational and clinical* **8**, p. 20552173221108635 (06 2022).
5. A. D'Costa, S. Denkovski, M. Malyska, S. Y. Moon, B. Rufino, Z. Yang, T. Killian and M. Ghassemi, Multiple sclerosis severity classification from clinical text, 7 (November 2020).
6. M. Davis, S. Sriram, W. Bush, J. Denny and J. Haines, Automated extraction of clinical traits of multiple sclerosis in electronic medical records, *Journal of the American Medical Informatics Association : JAMIA* **20** (10 2013).
7. E. Rouzic, M. Ghiani, E. Zhuleku, A. Dillenseger, U. Maywald, T. Wilke, T. Ziemssen and L. Craveiro, Claims-based algorithm to estimate the expanded disability status scale for multiple sclerosis in a german health insurance fund: a validation study using patient medical records, *Frontiers in Neurology* **14** (12 2023).
8. L. Kappos, Neurostatus scoring. slightly modified from jf kurtzke, *Neurology* **1983**, 1444 (2011).
9. J. F. Kurtzke, Historical and clinical perspectives of the expanded disability status scale, *Neuroepidemiology* **31**, 1 (2008).
10. U. Vurcan, S. Hwang, S. Thomas, A. Batugo, A. Acevedo, A. Kaviyarasu, A. BS, O. Mitchell and D. Mowery, Reliability in ai-assisted critical care: Assessing large language model robustness

and instruction following for cardiac arrest identification., in *NeurIPS 2024, the Thirty-Eighth Annual Conference on Neural Information Processing Systems Workshop*, 2024.
11. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models (2023).
12. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, Large language models are zero-shot reasoners (2023).

7. Appendix

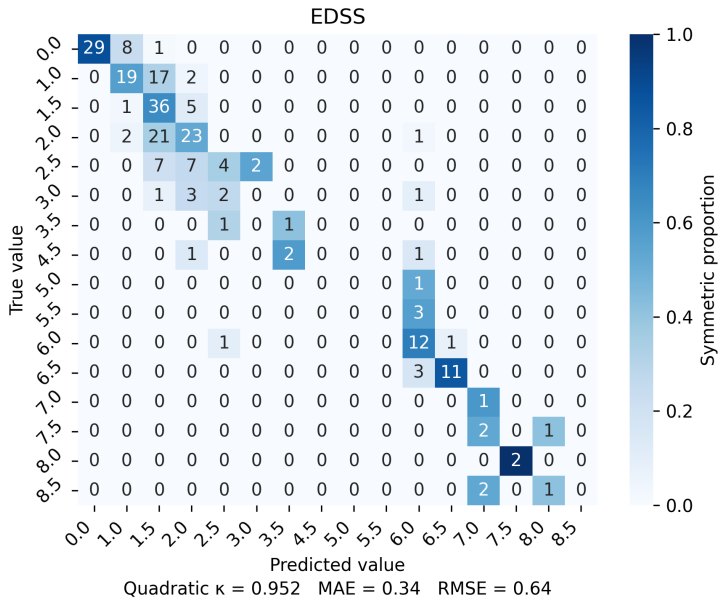

Fig. 4: EDSS confusion matrix

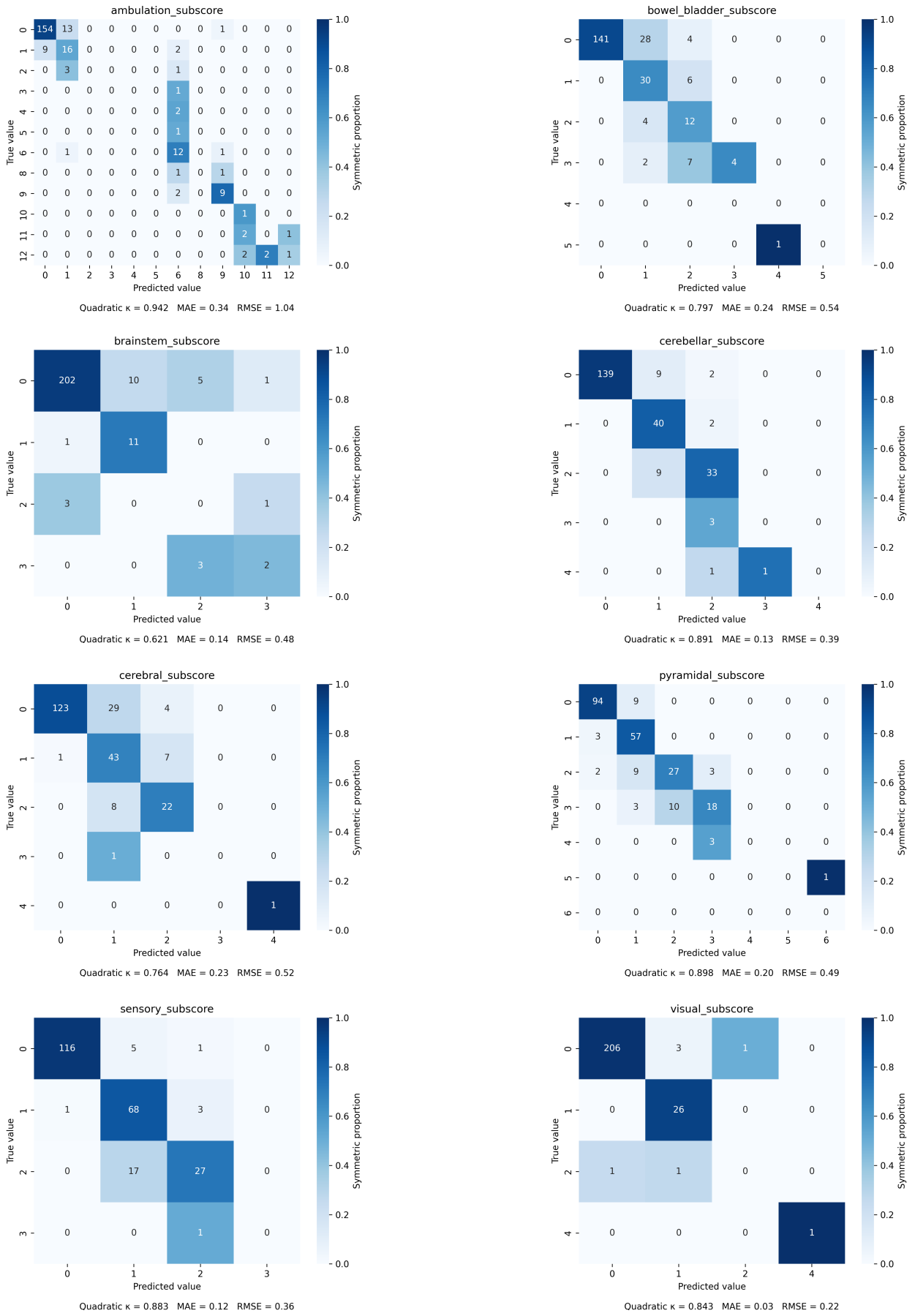

Fig. 5: Functional System (FS) confusion matrices for all subscore categories