Standardizing Physician Notes Improves Accuracy of Clinical Concept Extraction Without Information Loss

Daniel B. Hier *, Michael A. Carrithers, Steven Keith Platt, Anh Nguyen, Ioannis Giannopoulos, Tayo Obafemi-Ajayi

Posted Date: 18 April 2025

doi: 10.20944/preprints202504.1452.v1

Keywords: electronic health records; physician notes; standardization; human phenotype ontology; doc2hpo; large language models; data interoperability; concept extraction

Article

Standardizing Physician Notes Improves Accuracy of Clinical Concept Extraction Without Information Loss

Daniel B. Hier [1,*], Michael A. Carrithers [1], Steven K. Platt [2], Anh Nguyen [2], Ioannis Giannopoulos [2] and Tayo Obafemi-Ajayi [3]

[1] Dept. of Neurology & Rehabilitation, University of Illinois at Chicago, Chicago, IL, USA
[2] Laboratory for Applied Artificial Intelligence, Loyola University, Chicago, IL, USA
[3] Engineering Program, Missouri State University, Springfield, MO, USA
* Correspondence: dhier@uic.edu

Abstract: Clinician notes are a rich source of patient information but often contain inconsistencies due to varied writing styles, abbreviations, medical jargon, grammatical errors, and non-standard formatting. These inconsistencies hinder the meaningful extraction of data from electronic health records, limiting their utility for quality improvement, population health, precision medicine, decision support, and research. We present a large language model approach to standardizing a corpus of 1,618 clinical notes. Standardization corrected grammatical and spelling errors, converted non-standard terms to standardized terminology, and expanded abbreviations and acronyms. Notes were also restructured using consistent section headings. Expert review of randomly sampled notes found no significant information loss. The F1 score for Human Phenotype Ontology concept extraction increased from 0.40 to 0.61 following standardization, indicating improved accuracy without sacrificing clinical content. We conclude that standardizing physician notes enhances their usability for downstream applications such as clinical care, concept extraction, exploratory data analysis, and data exchange.

Keywords: electronic health records; physician notes; standardization; human phenotype ontology; doc2hpo; large language models; data interoperability; concept extraction

1. Introduction

Electronic Health Records (EHRs) have transformed healthcare documentation by improving the legibility and availability of patient data [1]. They solved the "availability problem"—where paper charts were often inaccessible—and the "legibility problem" of handwritten notes [2–4]. Yet, physician dissatisfaction with EHRs remains high, driven by poor interface design, excessive documentation burden, clerical overload, and limited perceived benefit for patient care [5–12].

Since the release of ChatGPT in late 2022, interest in large language models (LLMs) has surged in healthcare [13–17]. LLMs are now being explored for a range of clinical applications: decision support, diagnosis explanation, summarization, concept extraction, ontology mapping, and transforming documents into interoperable formats [18–23]. Among these, LLM-based clinical note standardization holds special promise for improving note quality and downstream reusability. Some recent surveys show increasing physician satisfaction with EHRs due to improvements referable to AI [24]/

We define *standardization* as the process of improving the structural and linguistic integrity of a clinical note without altering its clinical meaning. This includes correcting spelling and grammar, replacing colloquialisms with standardized terms (e.g., `upgoing toe` → `Babinski sign`), expanding abbreviations (e.g., `OU` → `both eyes`), and enforcing consistent formatting and section headers.

Clinical notes are typically written through direct entry, dictation, copy-paste, or auto-insertion of structured fields. As a result, they often suffer from inconsistent structure, misspellings, excessive abbreviations, slang, and ambiguous terminology [25–30]. Up to 20% of note content may be abbreviations or acronyms [31–38]. Moreover, many notes lack machine-readable codes (e.g., ICD-11, SNOMED CT, LOINC, RxNorm) and are poorly formatted or excessively verbose [39–42].

 doi:10.20944/preprints202504.1452.v1

LLMs can help address these limitations by:

1. Expanding abbreviations based on context (e.g., MS" → multiple sclerosis" or mental status").
2. Correcting typographical and grammatical errors to improve clarity [43].
3. Replacing colloquialisms and dictation artifacts with standardized terminology (e.g., heart attack" → "myocardial infarction") [44].
4. Structuring notes into canonical sections (e.g., *History*, *Examination*, *Impression*, *Plan*) to enhance navigability.

These improvements (Figure 1) can:

1. Enhance the readability and clinical utility of notes.
2. Improve the accuracy of pipelines that extract and normalize medical concepts to standard ontologies.
3. Facilitate conversion of clinical notes into structured formats for data exchange such as Fast Healthcare Interoperability Resources (FHIR) [45–48].

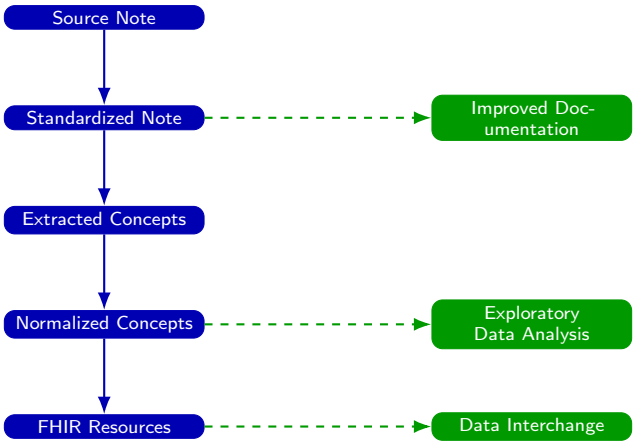

Figure 1. Potential Downstream Use Cases for Standardized Notes. Standardized notes offer opportunities for improved clinical documentation. Thes notes can be mined for medical concepts and then normalized to an appropriate ontology to support exploratory data analysis and machine learning and conversion to Fast Interoperability Healthare Resources for data exchange.

2. Methods

This study evaluates the effectiveness of GPT-4 in standardizing 1,618 de-identified clinical notes from a Neurology Clinic. We hypothesize that standardization will improve structural consistency, linguistic clarity, and the extractability of clinical concepts, thereby enhancing downstream applications such as ontology-based normalization, natural language processing (NLP), and data interoperability.

The evaluation framework (Figure 2) compares concept extraction accuracy under three conditions: from raw notes, from GPT-4-standardized notes, and from ground truth phrases. We define *concept extraction* as identifying relevant clinical terms or phrases in free text. *Concept normalization* refers to mapping each extracted term to a standardized ontology concept with its corresponding identifier. Notably, correct identification of a term does not guarantee accurate normalization, as models may capture semantic meaning without retrieving the correct machine-readable code [49–55].

 doi:10.20944/preprints202504.1452.v1

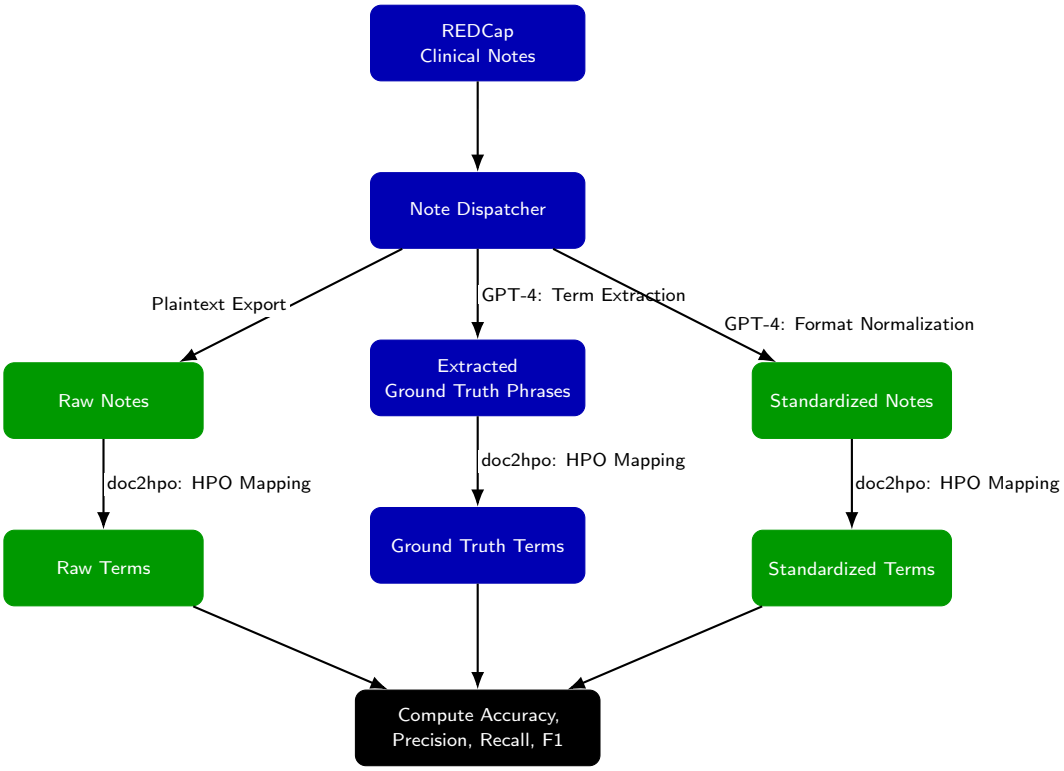

Figure 2. Experimental Setup for Evaluating the Accuracy of Note Standardization. Each clinical note is processed along three parallel paths: (1) raw notes are analyzed directly, (2) standardized notes are analyzed after GPT-4-based processing, and (3) ground truth phrases are analyzed after identification by GPT-4. In all paths, text is mapped to Human Phenotype Ontology (HPO) terms using *doc2hpo*, producing a set of terms from each path. Term sets from the raw and standardized notes are then compared to the ground truth set to compute accuracy, precision, recall, and F1 score.

Data Acquisition: Institutional Review Board (IRB) approval was obtained from the University of Illinois to use de-identified clinical notes from the Neurology Clinic. All notes were de-identified using REDCap [56]. We selected notes containing neuroimmunological diagnoses, including multiple sclerosis, myasthenia gravis, neuromyelitis optica, and Guillain–Barré syndrome. From an initial pool of 21,028 notes (2016–2022), a subset meeting the following inclusion criteria was selected: (1) outpatient visits, (2) authored in the Neurology Clinic, (3) written by physicians (residents or attendings), and (4) minimum note length of 2,000 characters. The final dataset comprised 1,618 typed notes exported as ASCII text and stored in a JSON-compatible format (Figure 3).

```
{
    "accession_num": "1",
    "note_text":
    "NEUROLOGY CLINIC NOTE
    Chief Complaint: New onset of double vision.
    History: History of optic neuritis and numbness.
    Examination: Increased reflexes. Babinski sign.
                 Internuclear ophthalmoplegia.
    Impression: Probable multiple sclerosis.
    Plan: MRI of brain.
          Start intravenous methylprednisolone."
}
```

Figure 3. Example neurology note converted to JSON format. Note has been truncated for brevity.

Note Standardization: GPT-4o (Open AI) was prompted to standardize neurology notes (Appendix A) into a standard format (Appendix B). GPT-4 produced a detailed list of all modifications for each note (Appendix C). Character count, word count, sentence count, spelling and grammar corrections,

 doi:10.20944/preprints202504.1452.v1

abbreviation expansion, and word substitutions were calculated for each note (Figure 5). To evaluate fidelity, 100 randomly selected note pairs were manually reviewed to ensure no clinically significant information was omitted and to screen for any spurious additions, such as hallucinations or confabulations.

Concept Extraction Using doc2hpo: We used the open-source Python implementation of doc2hpo [57], executed locally with the Aho-Corasick string-matching algorithm and a custom negation detection module ('NegationDetector') that applies a windowed keyword search within sentence boundaries. We used the HPO release dated 2025-03-03. No modifications were made to the *doc2hpo* source code. doc2hpo was applied under three conditions:

1. Raw notes (unmodified exports from REDCap),
2. Standardized notes (post-GPT-4 processing),
3. Ground truth phrases (GPT-4 extracted candidate HPO terms).

For the ground truth, GPT-4 was prompted to identify all candidate HPO-eligible phrases in each note (Appendix D). These phrases were combined into a single synthetic sentence for *doc2hpo* input (e.g., "The patient has ataxia, weakness, aphasia, and hyperreflexia."). *doc2hpo* returned a list of matched terms and corresponding HPO IDs. For each of the 1,618 notes, we compiled three term sets:

1. **Ground truth terms:** GPT-4 extracted phrases successfully matched by doc2hpo.
2. **Raw-note terms:** doc2hpo output from raw notes.
3. **Standardized-note terms:** doc2hpo output from standardized notes.

Metric Computation: For each extracted term, we classified its match status as follows:

1. **True Positive (TP):** Term present in both the extracted set and the ground truth set.
2. **False Negative (FN):** Ground truth term not captured in extraction set.
3. **False Positive (FP):** Term extracted but not part of ground truth set.
4. **True Negative (TN):** Term present in the note but not successfully mapped to HPO by doc2hpo.

Performance metrics (precision, recall, accuracy, and F1 score) were calculated by the method of Powers [58].

3. Results and Discussion

3.1. Note Standardization Improves Structure and Readability

We used GPT-4 to standardize 1,618 de-identified clinical notes from a neurology clinic. Each note was converted from plaintext to a structured JSON format with high-level section headers (e.g., *History*, *Examination*, *Impression*, *Plan*). GPT-4 was prompted to standardize each note by correcting spelling and grammatical errors, expanding acronyms and abbreviations, and replacing non-standard phrasing (Appendix E). GPT-4 corrected and average of 4.9 ± 1.8 grammatical errors, 3.3 ± 5.2 spelling errors, and expanded 12.6 ± 8.1 abbreviations or acronyms per note. Additionally, it substituted standardized clinical terminology for 3.1 ± 3.0 instances of jargon or non-standard language. A manual review of 100 randomly selected notes confirmed the preservation of clinical content, with improvements in clarity, structure, and overall readability. No hallucinations or spurious additions were identified during human expert review.

3.2. Standardization Improves HPO Term Extraction by doc2hpo

To evaluate whether GPT-4-based standardization improves phenotype term extraction, we compared three parallel data flows:

- **Raw Notes** → doc2hpo → Raw Terms
- **Standardized Notes** → doc2hpo → Standardized Terms
- **GPT-4 Extracted Phrases** → doc2hpo → Ground Truth Terms

All extracted terms were mapped to the HPO using *doc2hpo*. The ground truth was defined as HPO terms extracted by doc2hpo from GPT-4-supplied candidate phrases. For each note, term sets

 doi:10.20944/preprints202504.1452.v1

from the raw and standardized notes were evaluated against the corresponding ground truth set using standard classification metrics: precision, recall, accuracy, and F1 score.

GPT-4-standardized notes yielded higher precision and accuracy compared to raw notes. Recall also improved, rising from 0.61 to 0.68. Although modest in absolute terms, this gain was statistically significant due to the large sample size (n = 1,618) and contributed to a measurable improvement in F1 score (Figure 4). The combined improvements suggest that GPT-4 preprocessing enhances overall extraction performance without sacrificing sensitivity. The simultaneous increase in both precision and recall indicates that GPT-4 did not introduce spurious terms or omit clinically relevant information. These results support the hypothesis that physician note standardization by LLMs improves the performance of NLP pipelines performing ontology-term normalization.

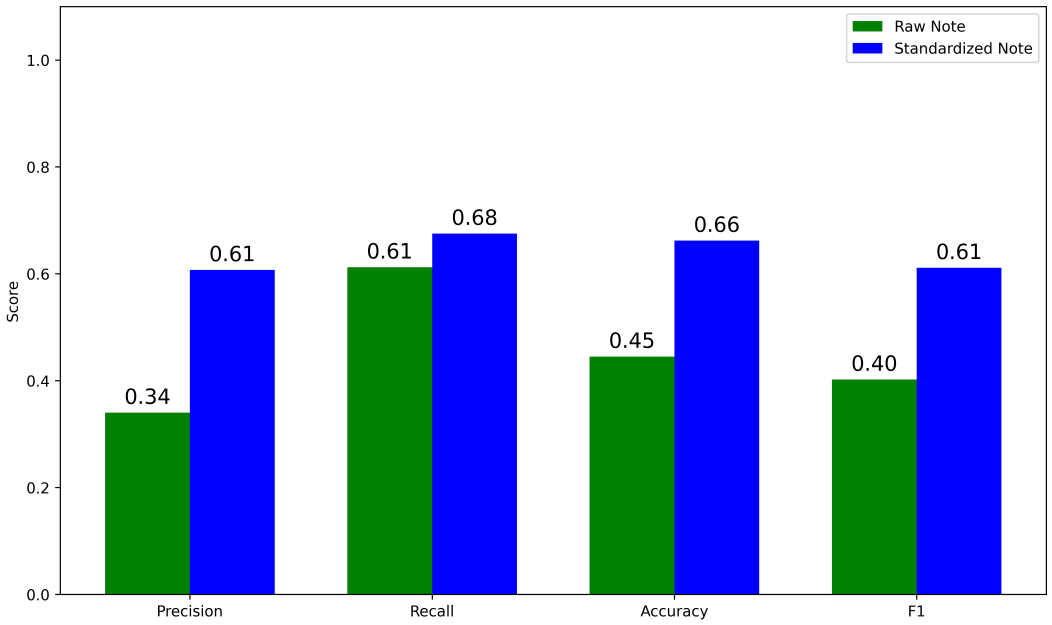

Figure 4. Performance metrics for doc2hpo on normalizing extracted terms from raw and standardized to their correct HPO term and HPO ID. Accuracy, precision, and recall improve after GPT-4-based note standardization ($p < 0.01$, n = 1592, two-way t-test.)

3.3. Downstream Use Cases After Note Standardization

Once physician notes are standardized by a large language model, several downstream applications become feasible (Figure 1):

1. Standardized notes can be integrated in real time into the EHR to enhance note quality, readability, and clinical interpretability.
2. Extracted terms can be passed into NLP pipelines that map these terms to ontology concepts, making clinical data computable for exploratory analysis, machine learning, population health, and quality improvement.
3. Normalized concepts can be restructured as FHIR resources to support efficient data exchange [45–48,59].

Major medical centers are already exploring the use of LLMs to analyze free text in the EHR, driving predictive analytics [60] and improving documentation quality among house officers [61].

 doi:10.20944/preprints202504.1452.v1

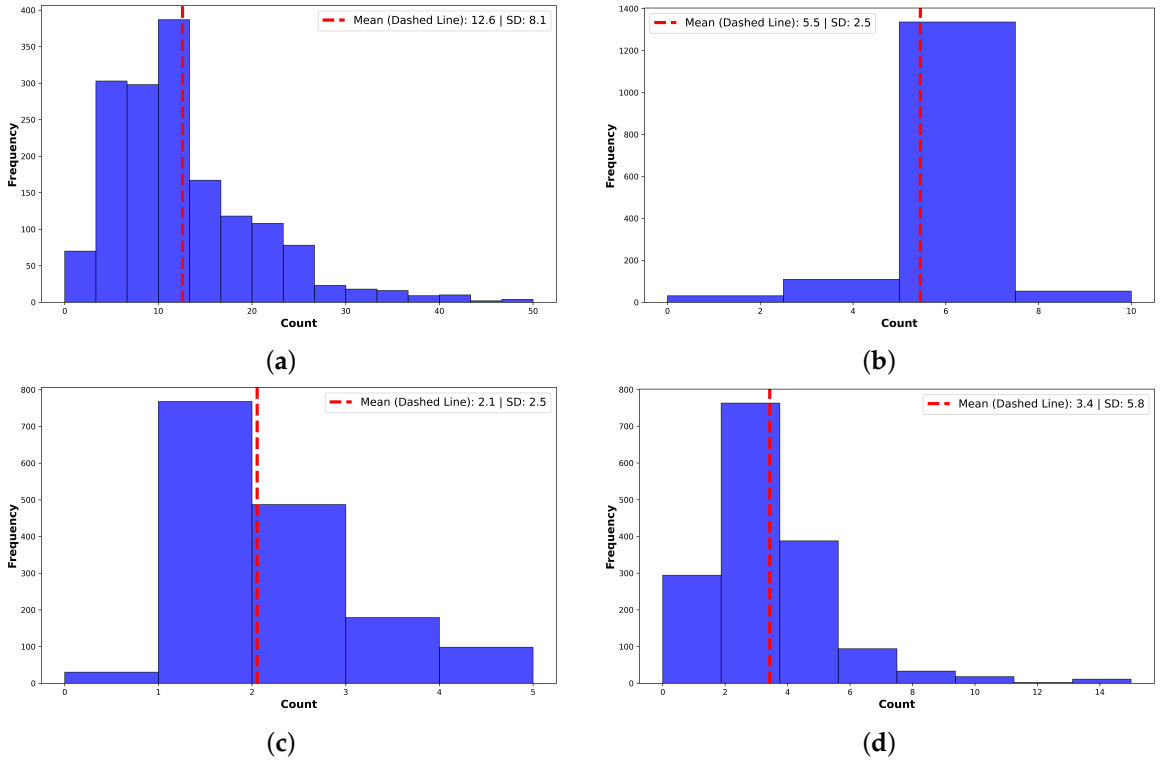

Figure 5. Summary of Corrections made by GPT-4o. (a) Acronyms and Abbreviations. Normalization expanded a mean of 12.6 ± 8.1 per note. (b) Grammatical Errors. Normalization corrected a mean of 4.9 ± 1.8 grammatical errors per note. (c) Slang, Jargon, and Non-Standard Terms. GPT-4 corrected 3.1 ± 3.0 non-standard terms per note. Example: "feeling blue" $\rightarrow$ "symptoms of depression". (d) Spelling Errors. GPT-4 corrected a mean of 3.3 ± 5.2 errors per note. High variance reflects many outlier notes.

3.4. Addressing Systemic Challenges in EHR Documentation

While large language models improve note structure and readability, systemic challenges in EHR documentation remain. Copy-and-paste practices, unverified carry-forward data, and documentation that omits or exaggerates care are beyond the scope of current AI tools [62–64]. Standardization alone does not reduce documentation burden, which is more fundamentally tied to EHR design, clinician workflows, and organizational practices [10,65–67]. Promising technologies like ambient AI may help, but broader reforms are needed to address who documents, what is documented, and how. Encouragingly, recent surveys suggest that the introduction of AI tools is already improving physician satisfaction with the EHR [24].

3.5. Limitations

This study analyzed 1,618 de-identified neurology notes, primarily from patients with multiple sclerosis. Broader validation across diverse diagnoses, note types, and clinical settings is needed. Ground truth medical concepts were identified using GPT-4 with human expert review. While this approach enabled scalable evaluation, future work should incorporate fully manual annotation for comparison. We did not assess the processing time or computational costs of standardization. Although deploying LLMs in production may incur costs, many institutions are actively exploring this path [61,68]. De-identification ensured HIPAA compliance, but explicit consent from clinicians and patients was not obtained. In active clinical environments, reconfiguration of notes may require additional safeguards or institutional approval. Nevertheless, our results support the growing interest in note reconfiguration, template engineering, and documentation quality. Once digitized, clinical text can be adapted to multiple formats, enhancing interoperability and downstream utility [61,69–72].

 doi:10.20944/preprints202504.1452.v1

4. Conclusions

Standardizing physician notes with large language models improves their readability, consistency, and interoperability without loss of clinical content. In this study, GPT-4 enhanced note structure, corrected errors, and expanded abbreviations, facilitating more accurate extraction and normalization of medical concepts by doc2hpo. These findings suggest that LLMs can augment both clinical and research workflows by making unstructured text more usable for downstream NLP and ontology-based analysis. As adoption grows, early evidence indicates that such tools may also contribute to improved physician satisfaction with electronic health records [24].

Appendix A. Prompt to GPT-4 to Standardize a Physician Note

```
You are a highly skilled medical terminologist
specializing in clinical note standardization.
Your task is to standardize the note using the following rules:

1. Expand abbreviations (e.g., BP → blood pressure),
   retaining common abbreviations in parentheses.

2. Correct spelling and grammar while preserving meaning.

3. Reorganize content under the following headings:
   History, Vital Signs, Examination, Labs, Radiology,
   Impression, and Plan.

4. Replace non-standard terms with standard clinical terminology.
```

Appendix B. Note Format Used by GPT-4 for Standardized Notes

```
{
  "HISTORY": {
     "Chief Complaint": "...",
     "Interim History": "..."
  },
  "VITAL SIGNS": {
     "Blood Pressure": "...",
     "Pulse": "...",
     "Temperature": "...",
     "Weight": "..."
  },
  "EXAMINATION": {
     "Mental Status": "...",
     "Cranial Nerves": "...",
     "Motor": "...",
     "Sensory": "...",
     "Reflexes": "...",
     "Coordination": "...",
     "Gait and Station": "..."
  },
  "LABS": "...",
  "RADIOLOGY": "...",
  "IMPRESSION": {
     "Assessment": "..."
  },
  "PLAN": {
     "Testing": "...",
     "Education Provided": {
        "Instructions": "...",
        "Barriers to Learning": "...",
        "Content": "...",
        "Outcome": "..."
     },
     "Return Visit": "..."
```

 doi:10.20944/preprints202504.1452.v1


```
  },
  "Metrics": {
     "Grammatical Errors": n,
     "Abbreviations and Acronyms Expanded": ["..."],
     "Spelling Errors": ["..."],
     "Non-Standard Terms Corrected": ["..."]
  }
}
```

Appendix C. Example of Corrections Made by GPT-4

```
"Abbreviations Expanded": [
  "BP", "IVIG", "MRI", "EMG", "PT", "OTC", "OT", "CSF",
  "WBC", "RBC", "HSV", "PCR", "CIDP", "INCAT", "BPD",
  "CBD", "BSA", "FPL", "EHL", "FN", "PA"
],
"Spelling Errors Corrected": [
  "wreight", "materal", "unknwon", "schizphernia", "tjhan"
],
"Non-Standard Terms Mapped": [
  "heart attack -> myocardial infarction"
]
```

Appendix D. Prompt to GPT-4 to Identify Ground Truth Terms

```
You are an expert medical coder with expertise in medical
terminologies such as the Human Phenotype Ontology (HPO).
From the note text {note_text}, extract all potential HPO terms
the patient may have. Return a JSON object with a list of
extracted terms under the key "hpo_terms".

Use this exact format:
{
  "hpo_terms": ["term1", "term2", "term3"]
}
```

Appendix E. Sample Neurological Examinations: Before and After Standardization

Neurological Examination Prior to Standardization. Expanded terms are highlighted in blue.

```
Neurologic:
Mental status : awake , alert , oriented to person ,
    place , and time.  Follows commands briskly ,
    including 2 step commands.  Naming and repetition intact.
    Fluent speech with no dysarthria.
Cranial nerves : PERRL , no rAPD ,
    unable to perform full L
    lateral gaze but otherwise EOMI , facial sensation full
    and symmetric , smile full and symmetric , palate and
    uvula elevate symmetrically , shoulder shrug intact ,
    tongue midline
...continues....
```

Neurological Examination After Standardization by GPT-4.

```
EXAMINATION:
    Mental Status: Awake, alert, oriented to person, place,
        and time. Follows commands briskly, including
        two-step commands. Naming and repetition intact.
        Fluent speech with no dysarthria.
    Cranial Nerves: Pupils equal, round, and reactive
        to light, no relative afferent pupillary
```

```
        defect, unable to perform full left lateral gaze but
        otherwise extraocular movements are intact,
        facial sensation full and symmetric,
        smile full and symmetric, palate and uvula elevate
        symmetrically, shoulder shrug intact, tongue midline
...continues....
```

Author Contributions: Conceptualization, All authors; Methodology, D.B.H. and T.O.; Software, D.B.H.; Validation, all authors; Formal analysis, D.B.H. and T.O.; Investigation and data acquisition, M.A.C.; Data curation, D.B.H. and M.A.C.; Writing—original draft preparation, D.B.H.; Writing—review and editing, all authors; Visualization, D.B.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science Foundation under Award Number 2423235.

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: Informed consent was obtained from all participants as part of enrollment in the UIC Neuroimmunology Biobank.

Data Availability Statement: Data and code supporting this study are available from the corresponding author on request.

Acknowledgments: We thank the UIC Neuroimmunology BioBank Team for their support.

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

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
EHR	Electronic Health Record
FHIR	Fast Healthcare Interoperability Resources
HIPAA	Health Insurance Portability and Accountability Act
HPO	Human Phenotype Ontology
ICD	International Classification of Diseases
IRB	Institutional Review Board
JSON	JavaScript Object Notation
LLM	Large Language Model
LOINC	Logical Observation Identifiers Names and Codes
NLP	Natural Language Processing
REDCap	Research Electronic Data Capture
RxNorm	Standardized Nomenclature for Clinical Drugs
TP/FN/FP/TN	True Positive / False Negative / False Positive / True Negative

References

1. Menachemi, N.; Collum, T.H. Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy* **2011**, pp. 47–55.
2. Bruner, A.; Kasdan, M.L. Handwriting errors: harmful, wasteful and preventable. *Journal-Kentucky Medical Association* **2001**, *99*, 189–192.
3. Kozak, E.A.; Dittus, R.S.; Smith, W.R.; Fitzgerald, J.F.; Langfeld, C.D. Deciphering the physician note. *Journal of General Internal Medicine* **1994**, *9*, 52–54.
4. Rodríguez-Vera, F.J.; Marin, Y.; Sanchez, A.; Borrachero, C.; Pujol, E. Illegible handwriting in medical records. *Journal of the Royal Society of Medicine* **2002**, *95*, 545–546.

 doi:10.20944/preprints202504.1452.v1

5. Holmgren, A.J.; Hendrix, N.; Maisel, N.; Everson, J.; Bazemore, A.; Rotenstein, L.; Phillips, R.L.; Adler-Milstein, J. Electronic health record usability, satisfaction, and burnout for family physicians. *JAMA Network Open* **2024**, *7*, e2426956–e2426956.
6. Muhiyaddin, R.; Elfadl, A.; Mohamed, E.; Shah, Z.; Alam, T.; Abd-Alrazaq, A.; Househ, M. Electronic health records and physician burnout: a scoping review. *Informatics and Technology in Clinical Care and Public Health* **2022**, pp. 481–484.
7. Downing, N.L.; Bates, D.W.; Longhurst, C.A. Physician burnout in the electronic health record era: are we ignoring the real cause? *Annals of Internal Medicine* **2018**, *169*, 50–51.
8. Elliott, M.; Padua, M.; Schwenk, T.L. Electronic health records, medical practice problems, and physician distress. *International Journal of Behavioral Medicine* **2022**, pp. 1–6.
9. Carroll, A.E. How health information technology is failing to achieve its full potential. *JAMA Pediatrics* **2015**, *169*, 201–202.
10. Rodríguez-Fernández, J.M.; Loeb, J.A.; Hier, D.B. It's time to change our documentation philosophy: writing better neurology notes without the burnout. *Frontiers in Digital Health* **2022**, *4*, 1063141.
11. Koopman, R.J.; Steege, L.M.B.; Moore, J.L.; Clarke, M.A.; Canfield, S.M.; Kim, M.S.; Belden, J.L. Physician information needs and electronic health records (EHRs): time to reengineer the clinic note. *The Journal of the American Board of Family Medicine* **2015**, *28*, 316–323.
12. Budd, J. Burnout related to electronic health record use in primary care. *Journal of primary care & community health* **2023**, *14*, 21501319231166921.
13. Sahoo, S.S.; Plasek, J.M.; Xu, H.; Uzuner, Ö.; Cohen, T.; Yetisgen, M.; Liu, H.; Meystre, S.; Wang, Y. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *Journal of the American Medical Informatics Association* **2024**, p. ocae074.
14. Yan, C.; Ong, H.H.; Grabowska, M.E.; Krantz, M.S.; Su, W.C.; Dickson, A.L.; Peterson, J.F.; Feng, Q.; Roden, D.M.; Stein, C.M.; et al. Large language models facilitate the generation of electronic health record phenotyping algorithms. *Journal of the American Medical Informatics Association* **2024**, p. ocae072.
15. Munzir, S.I.; Hier, D.B.; Carrithers, M.D. High Throughput Phenotyping of Physician Notes with Large Language and Hybrid NLP Models. *arXiv preprint arXiv:2403.05920* **2024**. accepted in International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC 2024), https://doi.org/10.48550/arXiv.2403.05920.
16. Omiye, J.A.; Gui, H.; Rezaei, S.J.; Zou, J.; Daneshjou, R. Large language models in medicine: the potentials and pitfalls. *arXiv preprint arXiv:2309.00087* **2023**.
17. Clusmann, J.; Kolbinger, F.R.; Muti, H.S.; Carrero, Z.I.; Eckardt, J.N.; Laleh, N.G.; Löffler, C.M.L.; Schwarzkopf, S.C.; Unger, M.; Veldhuizen, G.P.; et al. The future landscape of large language models in medicine. *Communications Medicine* **2023**, *3*, 141.
18. Li, Y.; Wang, H.; Yerebakan, H.; Shinagawa, Y.; Luo, Y. Enhancing Health Data Interoperability with Large Language Models: A FHIR Study. *arXiv preprint arXiv:2310.12989* **2023**.
19. Van Veen, D.; Van Uden, C.; Blankemeier, L.; Delbrouck, J.B.; Aali, A.; Bluethgen, C.; Pareek, A.; Polacin, M.; Reis, E.P.; Seehofnerova, A.; et al. Clinical text summarization: adapting large language models can outperform human experts. *Research Square* **2023**.
20. Tang, L.; Sun, Z.; Idnay, B.; Nestor, J.G.; Soroush, A.; Elias, P.A.; Xu, Z.; Ding, Y.; Durrett, G.; Rousseau, J.F.; et al. Evaluating large language models on medical evidence summarization. *NPJ digital medicine* **2023**, *6*, 158.
21. Zhou, W.; Bitterman, D.; Afshar, M.; Miller, T.A. Considerations for health care institutions training large language models on electronic health records. *arXiv preprint arXiv:2309.12339* **2023**.
22. Qiu, J.; Li, L.; Sun, J.; Peng, J.; Shi, P.; Zhang, R.; Dong, Y.; Lam, K.; Lo, F.P.W.; Xiao, B.; et al. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics* **2023**.
23. Wang, Y.; Zhao, Y.; Petzold, L. Are Large Language Models Ready for Healthcare. *A Comparative Study on Clinical Language Understanding. ArXiv, abs/2304.05368* **2023**.
24. Henderson, J. Fewer Physicians Consider Leaving Medicine, Survey Finds. *MedPage Today* **2025**. Accessed March 29, 2025.
25. Kugic, A.; Schulz, S.; Kreuzthaler, M. Disambiguation of acronyms in clinical narratives with large language models. *Journal of the American Medical Informatics Association* **2024**, *31*, 2040–2046.
26. Aronson, J.K. When I use a word... Medical slang: a taxonomy. *bmj* **2023**, *382*.

27. Lee, E.H.; Patel, J.P.; VI, A.H.F. Patient-centric medical notes: Identifying areas for improvement in the age of open medical records. *Patient Education and Counseling* **2017**, *100*, 1608–1611.
28. Castro, C.M.; Wilson, C.; Wang, F.; Schillinger, D. Babel babble: physicians' use of unclarified medical jargon with patients. *American Journal of Health Behavior* **2007**, *31*, S85–S95.
29. Pitt, M.B.; Hendrickson, M.A. Eradicating jargon-oblivion—a proposed classification system of medical jargon. *Journal of General Internal Medicine* **2020**, *35*, 1861–1864.
30. Workman, T.E.; Shao, Y.; Divita, G.; Zeng-Treitler, Q. An efficient prototype method to identify and correct misspellings in clinical text. *BMC Research Notes* **2019**, *12*, 1–5.
31. Hamiel, U.; Hecht, I.; Nemet, A.; Pe'er, L.; Man, V.; Hilely, A.; Achiron, A. Frequency, comprehension and attitudes of physicians towards abbreviations in the medical record. *Postgraduate Medical Journal* **2018**, *94*, 254–258.
32. Myers, J.S.; Gojraty, S.; Yang, W.; Linsky, A.; Airan-Javia, S.; Polomano, R.C. A randomized-controlled trial of computerized alerts to reduce unapproved medication abbreviation use. *Journal of the American Medical Informatics Association* **2011**, *18*, 17–23.
33. Horon, K.; Hayek, K.; Montgomery, C. Prohibited abbreviations: seeking to educate, not enforce. *The Canadian Journal of Hospital Pharmacy* **2012**, *65*, 294.
34. Cheung, S.; Hoi, S.; Fernandes, O.; Huh, J.; Kynicos, S.; Murphy, L.; Lowe, D. Audit on the use of dangerous abbreviations, symbols, and dose designations in paper compared to electronic medication orders: A multicenter study. *Annals of Pharmacotherapy* **2018**, *52*, 332–337.
35. Shultz, J.; Strosher, L.; Nathoo, S.N.; Manley, J. Avoiding potential medication errors associated with non-intuitive medication abbreviations. *The Canadian journal of hospital pharmacy* **2011**, *64*, 246.
36. Baker, D.E. Campaign to Eliminate Use of Error-Prone Abbreviations. *Hospital Pharmacy* **2006**, *41*, 809–810.
37. Association, A.H.; of Health-System Pharmacists, A.S.; et al. Medication safety issue brief. Eliminating dangerous abbreviations, acronyms and symbols. *Hospitals & Health Networks* **2005**, *79*, 41–42.
38. Hultman, G.M.; Marquard, J.L.; Lindemann, E.; Arsoniadis, E.; Pakhomov, S.; Melton, G.B. Challenges and opportunities to improve the clinician experience reviewing electronic progress notes. *Applied Clinical Informatics* **2019**, *10*, 446–453.
39. McDonald, C.J.; Huff, S.M.; Suico, J.G.; Hill, G.; Leavelle, D.; Aller, R.; Forrey, A.; Mercer, K.; DeMoor, G.; Hook, J.; et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical Chemistry* **2003**, *49*, 624–633.
40. Hanna, J.; Joseph, E.; Brochhausen, M.; Hogan, W.R. Building a drug ontology based on RxNorm and other sources. *Journal of Biomedical Semantics* **2013**, *4*, 1–9.
41. Zarei, J.; Golpira, R.; Hashemi, N.; Azadmanjir, Z.; Meidani, Z.; Vahedi, A.; Bakhshandeh, H.; Fakharian, E.; Sheikhtaheri, A. Comparison of the accuracy of inpatient morbidity coding with ICD-11 and ICD-10. *Health Information Management Journal* **2025**, *54*, 14–24.
42. Lee, D.; de Keizer, N.; Lau, F.; Cornet, R. Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association* **2014**, *21*, e11–e19.
43. Ficarra, B.J. Grammar and Medicine. *Archives of Surgery* **1981**, *116*, 251–252.
44. Goss, F.R.; Zhou, L.; Weiner, S.G. Incidence of speech recognition errors in the emergency department. *International Journal of Medical Informatics* **2016**, *93*, 70–73.
45. Bender, D.; Sartipi, K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In Proceedings of the Proceedings of the 26th IEEE international symposium on computer-based medical systems. IEEE, 2013, pp. 326–331.
46. Vorisek, C.N.; Lehne, M.; Klopfenstein, S.A.I.; Mayer, P.J.; Bartschke, A.; Haese, T.; Thun, S. Fast healthcare interoperability resources (FHIR) for interoperability in health research: systematic review. *JMIR medical informatics* **2022**, *10*, e35724.
47. Braunstein, M.L. *Health Informatics on FHIR: How HL7's New API is Transforming Healthcare*; Springer, 2018.
48. Benson, T.; Grieve, G. Principles of health interoperability. *Springer International* **2021**, pp. 21–40.
49. 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*.
50. Hu, J.; Bao, R.; Lin, Y.; Zhang, H.; Xiang, Y. Accurate medical named entity recognition through specialized NLP models. *arXiv preprint arXiv:2412.08255* **2024**.
51. Luo, Y.F.; Henry, S.; Wang, Y.; Shen, F.; Uzuner, O.; Rumshisky, A. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association* **2020**, *27*, 1529–e1.

52. Agrawal, M.; O'Connell, C.; Fatemi, Y.; Levy, A.; Sontag, D. Robust benchmarking for machine learning of clinical entity extraction. In Proceedings of the Machine Learning for Healthcare Conference. PMLR, 2020, pp. 928–949.
53. 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**, p. 103526.
54. Funk, C.; Baumgartner, W.; Garcia, B.; Roeder, C.; Bada, M.; Cohen, K.B.; Hunter, L.E.; Verspoor, K. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC bioinformatics* **2014**, *15*, 1–29.
55. Zheng, J.G.; Howsmon, D.; Zhang, B.; Hahn, J.; McGuinness, D.; Hendler, J.; Ji, H. Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making* **2015**, *15*, 1–9.
56. Harris, P.A.; Taylor, R.; Thielke, R.; Payne, J.; Gonzalez, N.; Conde, J.G. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics* **2009**, *42*, 377–381.
57. Liu, C.; Peres Kury, F.S.; Li, Z.; Ta, C.; Wang, K.; Weng, C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic Acids Research* **2019**, *47*, W566–W570.
58. Powers, D.M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* **2020**.
59. Tayefi, M.; Ngo, P.; Chomutare, T.; Dalianis, H.; Salvi, E.; Budrionis, A.; Godtliebsen, F. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics* **2021**, *13*, e1549.
60. Jiang, L.Y.; Liu, X.C.; Nejatian, N.P.; Nasir-Moin, M.; Wang, D.; Abidin, A.; Eaton, K.; Riina, H.A.; Laufer, I.; Punjabi, P.; et al. Health system-scale language models are all-purpose prediction engines. *Nature* **2023**, *619*, 357–362.
61. Feldman, J.; Hochman, K.A.; Guzman, B.V.; Goodman, A.; Weisstuch, J.; Testa, P. Scaling note quality assessment across an academic medical center with AI and GPT-4. *NEJM Catalyst Innovations in Care Delivery* **2024**, *5*, CAT–23.
62. Siegler, E.L.; Adelman, R. Copy and paste: a remediable hazard of electronic health records. *The American Journal of Medicine* **2009**, *122*, 495–496.
63. Weiner, S.J.; Wang, S.; Kelly, B.; Sharma, G.; Schwartz, A. How accurate is the medical record? A comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. *Journal of the American Medical Informatics Association* **2020**, *27*, 770–775.
64. Sharma, R.; Kostis, W.J.; Wilson, A.C.; Cosgrove, N.M.; Hassett, A.L.; Moreyra, A.E.; Delnevo, C.D.; Kostis, J.B. Questionable hospital chart documentation practices by physicians. *Journal of General Internal Medicine* **2008**, *23*, 1865–1870.
65. Bakken, S. Can informatics innovation help mitigate clinician burnout? *Journal of the American Medical Informatics Association* **2019**, *26*, 93–94.
66. Kapoor, M. Physician burnout in the electronic health record era. *Annals of Internal Medicine* **2019**, *170*, 216.
67. Kang, C.; Sarkar, N. Interventions to reduce electronic health record-related burnout: a systematic review. *Applied Clinical Informatics* **2023**.
68. Ji, Z.; Wei, Q.; Xu, H. Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings* **2020**, *2020*, 269.
69. Epstein, J.A.; Cofrancesco, J.; Beach, M.C.; Bertram, A.; Hedian, H.F.; Mixter, S.; Yeh, H.C.; Berkenblit, G. Effect of outpatient note templates on note quality: NOTE (Notation Optimization through Template Engineering) randomized clinical trial. *Journal of General Internal Medicine* **2021**, *36*, 580–584.
70. Savoy, A.; Frankel, R.; Weiner, M. Clinical thinking via electronic note templates: who benefits? *Journal of General Internal Medicine* **2021**, *36*, 577–579.
71. Ebbers, T.; Kool, R.B.; Smeele, L.E.; Dirven, R.; den Besten, C.A.; Karssemakers, L.H.; Verhoeven, T.; Herruer, J.M.; van den Broek, G.B.; Takes, R.P. The impact of structured and standardized documentation on documentation quality; a multicenter, retrospective study. *Journal of Medical Systems* **2022**, *46*, 46.
72. Burke, H.B.; Hoang, A.; Becher, D.; Fontelo, P.; Liu, F.; Stephens, M.; Pangaro, L.N.; Sessums, L.L.; O'Malley, P.; Baxi, N.S.; et al. QNOTE: an instrument for measuring the quality of EHR clinical notes. *Journal of the American Medical Informatics Association* **2014**, *21*, 910–916.

 doi:10.20944/preprints202504.1452.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.