Large Language Model Based Assessment of Clinical Reasoning Documentation in the Electronic Health Record Across Two Institutions

Verity Schaye, David DiTullio, Benedict Guzman, Scott Vennemeyer, Hanniel Shih, Ilan Reinstein, Danielle E Weber, Abbie Goodman, Danny T. Y. Wu, Daniel J. Sartori, Sally A. Santen, Larry Gruppen, Yindalon Aphinyanaphongs, Jesse Burk Rafel

Submitted to: Journal of Medical Internet Research
on: October 24, 2024

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Large Language Model Based Assessment of Clinical Reasoning Documentation in the Electronic Health Record Across Two Institutions

Verity Schaye[1, 2]; David DiTullio[1]; Benedict Guzman[3]; Scott Vennemeyer[4]; Hanniel Shih[4]; Ilan Reinstein[2]; Danielle E Weber[5, 6]; Abbie Goodman[7, 6]; Danny T. Y. Wu[4]; Daniel J. Sartori[1]; Sally A. Santen[8]; Larry Gruppen[9]; Yindalon Aphinyanaphongs[3]; Jesse Burk Rafel[1, 2]

[1]Department of Medicine, NYU Grossman School of Medicine New York US
[2]Institute for Innovations in Medical Education, NYU Grossman School of Medicine New York US
[3]Division of Applied AI Technologies, NYU Langone Health New York US
[4]Department of Biostatistics, Health informatics, and Data Sciences, University of Cincinnati College of Medicine Cincinnati US
[5]Division of Hospital Medicine, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine Cincinnati US
[6]Division of Hospital Medicine, Department of Internal Medicine, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine Cincinnati US
[7]Division of Hospital Medicine, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati Cincinnati US
[8]Department of Emergency Medicine, University of Cincinnati College of Medicine Cincinnati US
[9]Department of Learning Health Sciences, University of Michigan Medical School Ann Arbor US

Corresponding Author:
Verity Schaye
Department of Medicine, NYU Grossman School of Medicine
550 1st Avenue
New York
US

Abstract

Background: Clinical reasoning is an essential skill, yet physicians receive limited feedback. Artificial intelligence holds promise to fill this gap.

Objective: We report the development of both named entity recognition (NER), logic-based and large language model (LLM)-based assessments of CR documentation in the electronic health record (EHR) across two institutions.

Methods: Two note sets were retrieved from the EHR at each institution (NYU Grossman School of Medicine (NYU) and University of Cincinnati College of Medicine (UC)): 1) retrospective dataset comprised of internal medicine resident admission notes from July 2020-December 2021 (n=700 NYU notes, n=450 UC notes) and 2) prospective validation dataset from July 2023-December 2023 (n=155 NYU notes, n=92 UC notes). Using a validated human gold standard for assessment of CR documentation, the R-DEA tool, clinicians rated notes for D (differential diagnosis) and EA (explanation of reasoning) quality, each on 3-point scales (D0, D1, D2 and EA0, EA1, EA2). Model training occurred accordingly on the retrospective datasets: 1) NYU development of NER, logic-based model with validation at UC, 2) NYU fine tune training of LLM NYUTron (a BERT-like (Bidirectional Encoder Representation with Transformer) LLM with about 110 million parameters that has been pre-trained on 7.25 million clinical notes), 3) NYU fine tune training of LLM GatorTron (an open source LLM with 345 million parameters that was pre-trained on over 82 billion words of de-identified clinical text), 4) UC fine tune training of NYU fine-tuned GatorTron, and 5) UC fine tune training of GatorTron. The best performing models were validated with the prospective datasets and performance assessed with F1 scores for the NER, logic-based model and AUROC and AUPRC for the LLMs.

Results: At NYU, the NYUTron models were the best performing. The D0 and D2 models with an AUROC 0.87, AUPRC 0.79 and AUROC 0.89, AUPRC 0.86, respectively. The D1 model did not have sufficient performance for implementation. The EA0 and EA1 models also did not have adequate performance so the approach pivoted to create a binary EA2 model (i.e. EA2 vs not EA2) which had excellent performance with an AUROC 0.85 and AUPRC 0.80.

At UC, the NER, D-logic-based model was the best performing D model. The F1-scores for the D model on the UC dataset were

0.80, 0.74, and 0.80 for D0, D1, D2, respectively. The UC fine tuning of NYU fine-tuned GatorTron EA2 model had an AUROC 0.75 and AUPRC 0.69.

Conclusions: This is the first study to our knowledge to demonstrate the use of LLMs for assessment of CR documentation quality in the EHR across two institutions. Lessons learned can help promote implementation of these technologies across institutions with ranges of technical resources and enhance feedback on the essential skill of CR.

(JMIR Preprints 24/10/2024:67967)
DOI: https://doi.org/10.2196/preprints.67967

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title: Large Language Model Based Assessment of Clinical Reasoning Documentation in the Electronic Health Record Across Two Institutions

Authors:
Verity Schaye, MD, MHPE,[1,4] David DiTullio, MD, PhD,[1] Benedict Guzman, MS,[2] Scott Vennemeyer, BS,[3] Hanniel Shih, MS,[3] Ilan Reinstein, MS,[4] Danielle E. Weber MD, MEd,[5,6] Abbie Goodman MD,[5,6] Danny T.Y. Wu PhD, MSI,[3] Daniel J. Sartori, MD,[1,4] Sally A. Santen MD, PhD,[7] Larry D. Gruppen, PhD,[8] Yindalon Aphinyanaphongs, MD, PhD,[2] Jesse Burk-Rafel, MD, MRes[1,4]

1. Department of Medicine, NYU Grossman School of Medicine, New York, New York, USA
2. Division of Applied AI Technologies, NYU Langone Health, New York, New York, USA
3. Department of Biostatistics, Health informatics, and Data Sciences, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA
4. Institute for Innovations in Medical Education, NYU Grossman School of Medicine, New York, New York, USA
5. Division of Hospital Medicine, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA
6. Division of Hospital Medicine, Department of Internal Medicine, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA
7. Department of Emergency Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA
8. Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, Michigan, USA

Corresponding Author:
Verity Schaye, MD, MHPE
NYU Grossman School of Medicine
550 1st avenue, MS G 61
NY, NY, 10016
ORCID ID: 0000-0003-0816-4037
Email: verity.schaye@nyulangone.org
Twitter: @vschaye

Abstract

Background

Clinical reasoning is an essential skill, yet physicians receive limited feedback. Artificial intelligence holds promise to fill this gap. We report the development of both named entity recognition (NER), logic-based and large language model (LLM)-based assessments of CR documentation in the electronic health record (EHR) across two institutions.

Methods

Two note sets were retrieved from the EHR at each institution (NYU Grossman School of Medicine (NYU) and University of Cincinnati College of Medicine (UC)): 1) retrospective dataset comprised of internal medicine resident admission notes from July 2020-December 2021 (n=700 NYU notes, n=450 UC notes) and 2) prospective validation dataset from July 2023-December 2023 (n=155 NYU notes, n=92 UC notes). Using a validated human gold standard for assessment of CR documentation, the R-DEA tool, clinicians rated notes for D (differential diagnosis) and EA (explanation of reasoning) quality, each on 3-point scales (D0, D1, D2 and EA0, EA1, EA2). Model training occurred accordingly on the retrospective datasets: 1) NYU development of NER, logic-based model with validation at UC, 2) NYU fine tune training of LLM NYUTron (a BERT-like (Bidirectional Encoder Representation with Transformer) LLM with about 110 million parameters that has been pre-trained on 7.25 million clinical notes), 3) NYU fine tune training of LLM GatorTron (an open source LLM with 345 million parameters that was pre-trained on over 82 billion words of de-identified clinical text), 4) UC fine tune training of NYU fine-tuned GatorTron, and 5) UC fine tune training of GatorTron. The best performing models were validated with the prospective datasets and performance assessed with F1 scores for the NER, logic-based model and AUROC and AUPRC for the LLMs.

Results

At NYU, the NYUTron models were the best performing. The D0 and D2 models with an AUROC 0.87, AUPRC 0.79 and AUROC 0.89, AUPRC 0.86, respectively. The D1 model did not have

sufficient performance for implementation. The EA0 and EA1 models also did not have adequate performance so the approach pivoted to create a binary EA2 model (i.e. EA2 vs not EA2) which had excellent performance with an AUROC 0.85 and AUPRC 0.80.

At UC, the NER, D-logic-based model was the best performing D model. The F1-scores for the D model on the UC dataset were 0.80, 0.74, and 0.80 for D0, D1, D2, respectively. The UC fine tuning of NYU fine-tuned GatorTron EA2 model had an AUROC 0.75 and AUPRC 0.69.

Conclusion

This is the first study to our knowledge to demonstrate the use of LLMs for assessment of CR documentation quality in the EHR across two institutions. Lessons learned can help promote implementation of these technologies across institutions with ranges of technical resources and enhance feedback on the essential skill of CR.

Introduction

Clinical reasoning (CR) is a fundamental skill that requires incorporating vast amounts of information into a prioritized differential diagnosis and treatment plan and therefore crucial that trainees are given feedback to improve.[1] Documentation in the electronic health record (EHR) can provide this opportunity. Furthermore, poor documentation can reflect lack of refined CR, and has been hypothesized to be linked to diagnostic errors.[2-4] While there are established methods to provide feedback, feedback can be limited by variability between faculty and limited time.[5-7]

Machine learning (ML), natural language processing (NLP), and other artificial intelligence (AI) technologies have emerged as avenues to augment feedback.[8-12] NLP has been used to automate scoring of documentation in simulated scenarios.[13-16] AI-augmented assessment of CR documentation has also been implemented in clinical environments; we have previously reported on an NLP-based supervised ML model that provides feedback on internal medicine (IM) residents CR documentation. However, this model was developed using earlier technologies and only provides binary feedback.[17] Similarly, Feldman et al published the development of a supervised ML model that provides binary feedback on CR documentation (quality of prioritized differential in progress notes).[18]

The more recent advances of generative AI (GAI) and large language models (LLMs) have expanded the potential of AI as a powerful tool to augment feedback.[19,20] However, while there is a building body of literature demonstrating use of LLMs in CR tasks, the majority of these studies are with curated medical data at single institutions and do not focus on assessment of human reasoning but on performance of LLM reasoning as compared to humans.[21-24] Navigating the use of LLMs with EHR data (vs curated data) can be much more complicated. There are the challenges of accessing the right data from the chart, a higher burden of accuracy, privacy issues, and variability in EHRs.[25-27]

Additionally, initial studies have shown LLMs do not perform as well digesting the complexity of information in the EHR to make accurate diagnoses.[27] Overall, while the performance of LLMs on CR tasks is promising, it is far from sufficient to replace humans and it is essential that we continue to provide feedback on our learners' CR.[19,27,28] AI-based tools remain an important strategy to enhance the amount of feedback we provide.[12]

Here, we report on an expansion of our prior work and describe the development across two institutions of a named entity recognition (NER), logic-based assessment and LLM-based assessments of resident CR documentation in the EHR that predicts the quality of CR across two domains using a validated human rubric.

Methods

Setting and Study Population

We conducted this study at two institutions: NYU Grossman School of Medicine (NYU) and University of Cincinnati College of Medicine (UC). NYU is a northeastern academic medical center with multiple hospital sites; the NYU IM residency program has two resident populations with separate recruitment processes and clinical rotations at different hospitals which use the same EHR – NYU Langone Health Manhattan and NYU Langone Health Brooklyn. NYU residents also rotate at two other sites with distinct EHRs not included in this study. In terms of technical resources, NYU has the infrastructure of both the Institute for Innovations in Medical Education which is a multidisciplinary team of clinician educators, data scientists, informaticians, and developers who apply the science of education and informatics to transform teaching, learning, evaluation, and assessment at NYU and the Division of Applied AI Technologies which focuses on using data and modeling to predict health outcomes across NYU Langone Health.[29,30] Additionally, both education and EHR data is stored and easily accessible via a central education data warehouse and there is access to a distributed-memory, high-performing computing cluster.[31,32]

UC is a midwestern academic medical center within which IM residents rotate at University of Cincinnati Medical Center (UCMC) and the Veterans Affairs Medical Center; only notes written at UCMC were included in the study. In terms of technical resources, UC has the Department of Biostatistics, Health Informatics, & Data Sciences (BHIDS) that enables the UC academic healthcare enterprise to make better use of biomedical data and technology for new discoveries, innovative science, and improved health care.[33] Although UC has access to many data sources across the health system, medical school and other education programs, there is not currently a centralized database for education and EHR data like at NYU. Additionally, UC has some access to high-performing computing resources, but NYU has a more developed infrastructure for using these tools for both clinical and educational use than at UC.

At each site two note sets were retrieved from an integrated EHR (Epic Systems, Verona, WI): 1) retrospective dataset comprised of IM resident admission notes from July 2020-December 2021 (n=700 NYU notes, n=450 UC notes) and 2) prospective validation dataset from July 2023-December 2023 (n=155 NYU notes, n=92 UC notes). The datasets at NYU were larger because the initial plan was for primary development and model fine tuning occurring at this institution.

Human Note Rating

We used the DEA components of the R-IDEA tool as our human rating gold standard.[6] We maintained the original D (differential diagnosis) score whether a note has an explicitly prioritized differential diagnosis with specific diagnoses (e.g. not diagnostic categories such as cardiac), scored as D0, D1, or D2. We discovered early in training experiments that discerning the E (explanation of lead diagnosis) and A (alternative diagnosis explained) would be difficult with available AI models, and iterated to create an overall *explanation of reasoning* EA score combining the E and A components (i.e., EA0, EA1, or EA2) (Figure 1). Six faculty (clinician educators with expertise in CR, assessment, and psychometrics), one resident, and one medical student reviewed admission notes to create the new EA score.

On the NYU retrospective dataset, raters also annotated spans of text using Prodigy (an annotation tool for creating training data for ML models). Annotations included five entity types for NER: three components of the D score (diagnosis (Dx), diagnostic category (DC), prioritization of diagnosis language (Prior)) and two components of the EA score (data (Data) and linkage terms (Link)).

To demonstrate interrater reliability, two raters labeled 76 notes from the NYU retrospective set for NER, D, and EA scores. For the UC notes, one rater from NYU and two raters from UC rated 20 notes for D and EA scores. Intraclass correlation (ICC) using a two-way ANOVA with mixed effects was calculated to assess rater consistency. The remainder of the notes at each institution were rated by one rater for D and EA scores. The remainder of the retrospective set at NYU was also rated by one rater for NER.

Note Preprocessing

We aimed to isolate the section of the assessment that concentrates on the differential diagnosis and the explanation of reasoning for the primary presenting problem. We iterated on prior truncation strategies outlined by Schaye et al[17] and different approaches were required at each institution given different note writing styles (details in the Supplementary Methods).

Model Development

Model training occurred with several approaches: 1) NYU development of NER, logic-based model with subsequent validation at UC, 2) NYU fine tune training of LLM NYUTron[34], 3) NYU fine tune training of LLM GatorTron[35], 4) UC fine tune training of NYU fine-tuned GatorTron, and 5) UC fine tune training of GatorTron. We were not able to validate NYUTron at UC pending contract execution for data sharing and will ideally do so in the future.

NER, Logic-Based Model Approach

We used a large, NLP word embedding model trained on scientific texts from the scispaCy library (en_core_sci_lg) with more than 700k vocabulary and 600k word vectors. We adjusted model weights with backpropagation using the human-annotated labels of the five entity types (Dx, DC, Prior, Data, Link).[36]

We calculated the predicted D scores (i.e. D0, D1, D2) using logic-based relationships between the extracted entities and the rating scale: D0, fewer than 2 unique diagnoses (Dx entity counts); D2, 2 unique diagnoses (Dx entity counts) and explicit prioritization (Prior entity counts); D1, everything else. We attempted ML models to predict EA score from the named entities, however, due to poor model performance, we abandoned further attempts to develop NER, logic-based EA scores.

In order to provide an impartial and dependable evaluation of the model's performance, we used ten-fold cross-validation. We calculated the Type NER (which demands some overlap between the system tagged entity and the gold standard annotation) evaluation metric, as suggested in SemEval-2013 Task 9, at each k-fold (details in the Supplementary Methods).[37] We report F1-score over ten-fold runs for each D score entity type (Dx, DC, and Prior) and D score prediction. The F1-score is crucial for evaluating NER-based models because it balances precision and recall, providing a comprehensive measure of a model's ability to correctly identify entities while minimizing both false positives and false negatives. This balance is essential for handling the often imbalanced nature of entity distributions in text, ensuring a more accurate assessment of model performance. We shared the best performing NER, logic-based D model with UC via a docker container and validated the model on the UC retrospective set.

LLM Approaches

NYUTron, developed by NYU, is a BERT-like (Bidirectional Encoder Representation with Transformer) LLM with about 110 million parameters that has been pre-trained on 7.25 million clinical notes (4.1 billion words, notes through May 2020).[34] We fine-tuned the model to classify D and EA scores. We applied a one-versus-rest approach, which resulted in the development and testing of six distinct models, each corresponding to a different D and EA score category (i.e., D0, D1, D2, EA0, EA1, and EA2 models). However, EA0 and EA1 models did not have adequate performance so we pivoted our approach to create a single binary EA2 model (i.e. EA2 vs not EA2). To evaluate model performance, we employed ten-fold cross-validation, with AUROC and AUPRC averaged over the ten runs. We chose AUROC and AUPRC as our primary metrics for all LLM-based models because these metrics are favored over F-scores for binary classification tasks, particularly with imbalanced datasets. They assess model performance across all possible thresholds, offering a more detailed understanding of trade-offs between true positives, false positives, and precision-recall dynamics, thereby aiding in the identification of the optimal decision-making threshold, which a single F-score cannot provide.

Unlike NYUTron, GatorTron is an open source LLM with 345 million parameters that was pre-trained on over 82 billion words of de-identified clinical text.[35] To enhance generalizability using an open source LLM, the same experiments described above for NYUTron were taken with GatorTron at NYU.

The NYU fine-tuned GatorTron EA2 model was shared with UC and further fine-tuned following a similar process. Due to the smaller set of notes and hardware limitations, particularly a relatively small Video Random Access Memory (VRAM) size of 16GB, some modifications were applied. A runtime text augmentation was implemented during training with the following settings: 15 words of synonym replacement, random word insertion, and random swap each, and finally, a random word deletion of 15% probability. Lastly, we applied random minority oversampling using inverse class frequency during training.[38] Additionally, given these limitations, we did not attempt to fine-tune the three separate NYU fine-tuned GatorTron D models at UC. Instead, we fine-tuned the original GatorTron model to predict all three possible D Scores with a single model at UC using the same training process and hyperparameters as the EA2 model. Further details on model hyperparameters and packages used at NYU and UC can be found in the Supplementary Methods.

Prospective Validation

As a final step of validation, we ran each of the best performing models selected for implementation on the site's prospective validation sets and assessed performance using F1-score for the NER, logic-based model and AUROC and AUPRC for the LLM models.

The study was approved by the NYU and UC institutional review boards.

RESULTS

Human Note Rating

At NYU, ICC was 0.83 (95% CI 0.74-0.89) and 0.77 (95% CI 0.65-0.85) for the D and EA scores, respectively, indicating substantial interrater agreement. Interannotator agreement across all 5 entity types averaged F1 score=0.81 (range 0.71-0.87 by entity type), indicating strong annotator overlap

(Figure 2). At UC, ICC was 0.83 (95% CI 0.68-0.92) and 0.84 (95% CI 0.70-0.93) for the D and EA scores, respectively.

In the both datasets at each institution there was a range of human-rated D and EA scores, diagnoses, and patient demographics (Table 1).

Model Performance

NER, Logic-Based Model

In the NYU retrospective dataset, the NER F1-score for the entity types used to compute the D score (Dx, DC, Prior) was 0.66. The NER model performed the best in extracting Prior and Dx entities with an F1-score of 0.75 and 0.68, respectively, but struggled with DC entities, achieving a 0.37 F1-score.

The NER, D logic-based model performed well at both sites with F1-scores of 0.83, 0.78, and 0.75 for D0, D1, and D2 scores, respectively at NYU and F1-scores of 0.75, 0.71, 0.76 for D0, D1, D2 scores, respectively at UC. At UC, the NER, D-logic-based model was the best performing D model overall selected for implementation and run on the UC prospective validation set with F1-scores of 0.80, 0.74, and 0.80 for D0, D1, D2 scores, respectively.

LLM-based Models

At NYU, NYUTron overall had better D and EA model performance on the retrospective set than GatorTron and were the best performing models overall (Table 2). However, while the D0 and D2 NYUTron models performed well, the D1 model was not performant on the retrospective set (AUROC 0.57 (CI 0.53-0.69), AUPRC 0.33 (CI 0.26-0.43)) and therefore was not suitable for implementation. As such, a stepwise approach was taken for the D1 model by taking advantage of the more performant D0 and D2 models (Table 2). The D0 and D2 NYUTron models had excellent performance on the prospective dataset as follows: D0 model, AUROC 0.87 and AUPRC 0.79 and D2 model, AUROC 0.89 and AUPRC 0.86 (Figure 3a).

Both the NYUTron EA0 and EA1 models had insufficient performance for implementation therefore the approach pivoted to create a single binary EA model – EA2 vs not EA2 (i.e., EA0 or EA1) (Table 2). The binary NYUTron EA2 model achieved sufficient performance for implementation with an AUROC 0.85 and AUPRC 0.80 on the prospective dataset (Figure 3b).

At UC, the NER, D logic model performed better than the D GatorTron models and were the D models implemented as described above (Table 2). The GatorTron EA2 model did reach sufficient performance for subsequent prospective validation with an AUROC 0.75 and AUPRC 0.69 (Table 2, Figure 3c).

A final step in optimizing performance was selecting thresholds for all LLM models implemented (details in the Supplementary Results).

Discussion

We developed both NER, logic-based and LLM-based assessments of CR documentation in the EHR across two institutions with different residency training cultures, expectations for documentation, and technical resources. This builds upon prior work of Schaye et al's supervised ML model to assess CR documentation, generating more specific feedback across two domains of the R-DEA tool.[17] We developed high performing D models that can provide feedback on a three-point scale and an EA model that can provide feedback on a two-point scale. To our knowledge this is the first study to apply LLMs to human CR in EHR data across institutions (rather than LLM reasoning on curated medical data).[17,18,21-24,27] Furthermore, despite the advances in LLMs, these technologies are not yet performant to replace human reasoning. AI-based tools such as the ones we developed can help ensure we are continuing to give our trainees feedback on the essential human skill of CR.[12,19,27,28] Next steps are to implement feedback generated from these AI-based assessments. We will iterate on dashboards implemented with our prior supervised ML model.[17] After implementation, we will collect data on impact on CR documentation practices of IM residents.

While we were able to navigate successfully some of the challenges of working with LLMs and EHR data such as accessing the right data from the chart and privacy issues,[25-27] we were not able to achieve sufficient performance of all the models at both sites. At NYU, we hypothesize that NYUTron performed better than GatorTron as NYUTron was developed on NYU EHR data.[34] Although, this also could lead to the potential of overfitting of the model. At UC we were not able to obtain adequate performance on the D GatorTron models but were able to with the D NER, logic-based model demonstrating that smaller NLP models can sometimes perform better than BERT models for specific tasks. It might not always be the newest technology needed to solve the task at hand and comparison of performance of different technologies can be a helpful strategy. Of note when this work initially began, GAI models such as ChatGPT were not readily available in HIPPA compliant instances at either institution but will be technology we integrate into future work.

We also learned a lot of lessons working across two institutions on how AI technologies can be adapted and successfully implemented at sites with different resources. Some key takeaways include: 1) experimenting with different LLMs including ones that are openly available[34,35]; 2) performing primary development at an institution with more resources and creating a HIPAA-compliant pipeline to share code 3) developing variations on truncation methods to account for different note writing styles; and 4) creating adaptable approaches to different degrees of computing power such as using text augmentation to prevent overfitting at UC.[38] We will take these lessons learned about generalizability to the next phases of the work and develop strategies to implement more advanced technologies across institutions while maintaining HIPAA compliance working with EHR data.

Limitations

While the models developed have high performance, they are not perfect. However, the intent is for use in formative and not high-stakes summative assessment which would have a higher threshold for implementation.[39] Another limitation is the LLMs used in this study only give a prediction of D and EA scores without explainability. As noted above, we will experiment further with newer GAI models in next phases of the work that can help address this limitation. Lastly, like our prior supervised ML model, these models only assess documentation of CR and not accuracy. Next phases of this work will include exploring accuracy and creating AI-based diagnostic performance feedback systems using newer GAI technologies.[40]

Conclusions

This is the first study to our knowledge to demonstrate the use of LLMs for assessment of CR documentation quality in the EHR across two institutions. Lessons learned from this study can help promote implementation of these technologies across institutions with ranges of technical resources. Further use of LLMs in the EHR for assessment and feedback can be transformative for medical education and patient care.

Figure 1: Explanation of Reasoning Score (Integrating the E score and A score from the original Revised-IDEA rubric**)***

Explanation of Reasoning:

Explains the reasoning behind the differential diagnosis, including supporting epidemiology and key features. There is clear linkage of supporting data to the diagnoses in the differential.

0: No explanation of reasoning

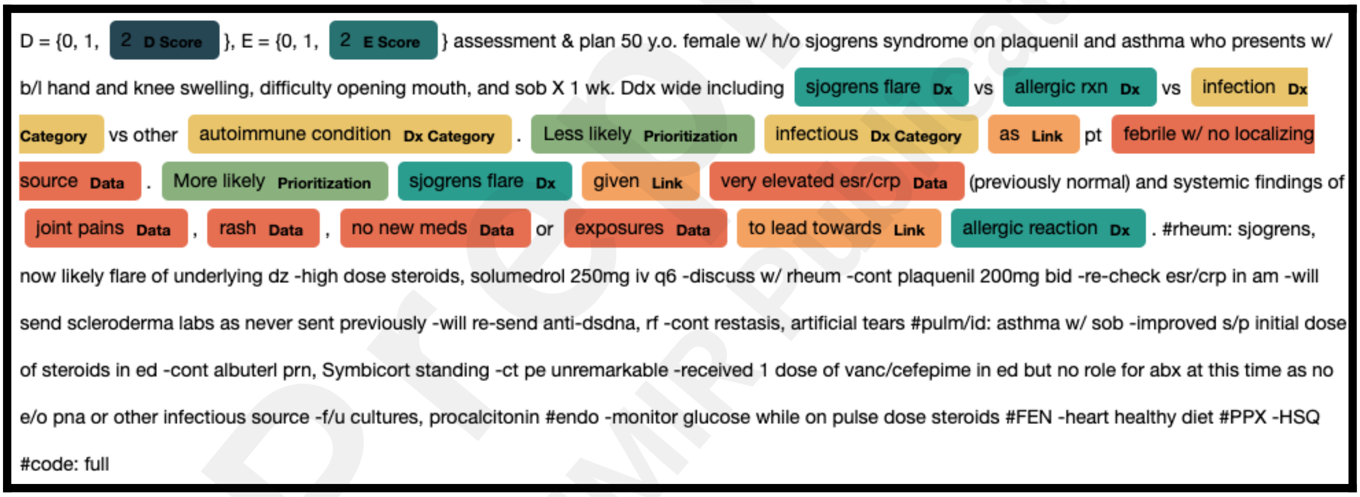
1: Explanation of reasoning includes at least 1 data point clearly linked to at least one diagnosis/diagnostic category on the differential

2: Explanation of reasoning includes at least 1 data point clearly linked to each of at least two of the diagnoses/diagnostic categories on the differential

*D score remained unchanged from the original Revised-IDEA rubric[6]

Figure 2. Example Note With Prodigy Labeling of D and EA Scores and Named Entity Recognition*

(Note has been modified to protect patient privacy)



D = {0, 1, 2 D Score }, E = {0, 1, 2 E Score } assessment & plan 50 y.o. female w/ h/o sjogrens syndrome on plaquenil and asthma who presents w/ b/l hand and knee swelling, difficulty opening mouth, and sob X 1 wk. Ddx wide including sjogrens flare Dx vs allergic rxn Dx vs infection Dx Category vs other autoimmune condition Dx Category . Less likely Prioritization infectious Dx Category as Link pt febrile w/ no localizing source Data . More likely Prioritization sjogrens flare Dx given Link very elevated esr/crp Data (previously normal) and systemic findings of joint pains Data , rash Data , no new meds Data or exposures Data to lead towards Link allergic reaction Dx . #rheum: sjogrens, now likely flare of underlying dz -high dose steroids, solumedrol 250mg iv q6 -discuss w/ rheum -cont plaquenil 200mg bid -re-check esr/crp in am -will send scleroderma labs as never sent previously -will re-send anti-dsdna, rf -cont restasis, artificial tears #pulm/id: asthma w/ sob -improved s/p initial dose of steroids in ed -cont albuterl prn, Symbicort standing -ct pe unremarkable -received 1 dose of vanc/cefepime in ed but no role for abx at this time as no e/o pna or other infectious source -f/u cultures, procalcitonin #endo -monitor glucose while on pulse dose steroids #FEN -heart healthy diet #PPX -HSQ

#code: full

*Overall F1-score for NER annotation of the 76 notes rated by two NYU raters was F1=0.81. NER interannotator agreement by entity type was as follows: F1-score for Diagnosis (Dx)=0.87, for Diagnostic Category (DC)=0.74, for Prioritization of Diagnosis Language (Prior)=0.87, for Data=0.80, and for Linkage Terms (Link)=0.71.

Table 1. Descriptive Statistics of Human-Rated Note Quality and Patient Characteristics

		NYU Retrospective Note Set (n=700)	NYU Prospective Note Set (n=155)	UC Retrospective Note Set (n=450)	UC Prospective Note Set (n = 92)
D Score n (%)	**0**	109 (15.6)	55 (35.5)	120 (26.7)	17 (18.5)
	1	154 (22.0)	46 (29.7)	155 (34.4)	31 (33.7)
	2	437 (62.4)	54 (34.8)	175 (38.9)	44 (47.8)
E Score n (%)	**0**	73 (10.4)	19 (12.3)	96 (21.3)	27 (29.3)
	1	255 (36.4)	74 (47.7)	171 (38.0)	21 (22.8)
	2	372 (53.1)	62 (40.0)	183 (40.7)	43 (46.7)
Patient age,y n (%)	**<= 54**	142 (20.3)	29 (18.7)	147 (32.7)	25 (27.2)
	55 - 68	181 (25.9)	35 (22.6)	184 (40.9)	37 (40.2)
	69 - 80	236 (33.7)	45 (29.0)	86 (19.1)	20 (21.7)
	>= 81	141 (20.1)	33 (21.3)	33 (7.3)	10 (10.9)
	NA	0 (0.0)	13 (8.4)	0 (0)	0 (0)
Patient Sex n (%)	**Female**	340 (48.6)	67 (43.2)	215 (47.8)	52 (56.5)
	Male	360 (51.4)	75 (48.4)	234 (52.0)	40 (43.5)
	NA	0 (0.0)	13 (8.4)	1 (0.2)	0 (0)
Primary Diagnosis by ICD-10, n (%)	**Cardiac**	134 (19.1)	38 (24.5)	42 (8.8)	12 (12.6)
	Dermatologic	13 (1.9)	3 (1.9)	18 (3.8)	4 (4.2)
	Endocrine	26 (3.7)	7 (4.5)	27 (5.7)	3 (3.2)
	Gastrointestinal	68 (9.7)	15 (9.7)	56 (11.8)	8 (8.4)
	Genitourinary	49 (7.0)	14 (9.0)	39 (8.2)	10 (10.5)
	Hematologic/ Oncologic	61 (8.7)	7 (4.5)	35 (7.4)	5 (5.3)
	Infectious	117 (16.7)	21 (13.5)	10 (2.1)	2 (2.1)
	Musculoskeletal	19 (2.7)	3 (1.9)	17 (3.6)	7 (7.4)
	Neurologic	21 (3.0)	1 (0.6)	12 (2.5)	4 (4.2)
	Other	54 (7.7)	11 (7.1)	149 (31.4)	30 (31.6)
	Psychiatric	5 (0.7)	7 (4.5)	13 (2.7)	2 (2.1)
	Pulmonary	72 (10.3)	10 (6.5)	47 (9.9)	7 (7.4)
	NA	61 (8.7)	18 (11.6)	10 (2.1)	1 (1.1)

Table 2. Large Language Model (LLM) Performance On Retrospective Note Sets For All NYUTron and GatorTron Experiments

LLM	Site	D/EA Score Classification	AUROC Retrospective Validation (CI)	AUPRC Retrospective Validation (CI)
NYUTron	NYU	D0	0.91 (0.85-0.93)	0.72 (0.58-0.76)

NYUTron	NYU	D1	0.57 (0.53-0.69)	0.33 (0.26-0.43)
NYUTron	NYU	D2	0.81 (0.80-0.87)	0.89 (0.85-0.93)
NYUTron	NYU	D1 Stepwise Approach*	N/A	N/A
NYUTron	NYU	EA0	0.83 (0.72-0.86)	0.36 (0.23-0.47)
NYUTron	NYU	EA1	0.74 (0.67-0.78)	0.63 (0.54-0.68)
NYUTron	NYU	EA2	0.84 (0.80-0.87)	0.84 (0.81-0.89)
NYUTron	NYU	EA2 Binary Model**	0.84 (0.81-0.85)	0.82 (0.80-0.87)
GatorTron	NYU	D0	0.92 (0.84-0.94)	0.72 (0.48-0.75)
GatorTron	NYU	D1	0.54 (0.5-0.59)	0.31(0.24-0.37)
GatorTron	NYU	D2	0.73 (0.78-0.85)	0.80 (0.82-0.92)
GatorTron	NYU	D1 Stepwise Approach*	N/A	N/A
GatorTron	UC	D0	0.75 (0.54-0.96)	0.51 (0.23-0.79)
GatorTron	UC	D1	0.61 (0.44-0.78)	0.46 (0.22-0.70)
GatorTron	UC	D2	0.72 (0.61-0.83)	0.63 (0.46-0.79)
GatorTron	NYU	EA0	0.80 (0.73-0.89)	0.42 (0.25-0.53)
GatorTron	NYU	EA1	0.75 (0.62-0.78)	0.63 (0.48-0.67)
GatorTron	NYU	EA2	0.83 (0.76-0.87)	0.83 (0.79-0.90)
GatorTron	NYU	EA2 Binary Model**	0.81 (0.76-0.87)	0.80 (0.79-0.90)
GatorTron	UC	EA0	N/A	N/A
GatorTron	UC	EA1	N/A	N/A
GatorTron	UC	EA2	N/A	N/A
GatorTron	UC	EA2 Binary Model**	0.72 (0.51-0.93)	0.63 (0.41-0.85)

*The D1 model ultimately did not have sufficient performance for implementation while the D0 and D2 had excellent performance so a stepwise approach was taken for the D1 score:

1) If D0 model predicts D=0, then D=0
2) If D2 model predicts D=2, then D=2
3) Else D=1

The NYU D1 Stepwise Approach achieved precision of 0.79 while the GatorTron D1 Stepwise Approach achieved precision of 0.73.

**Both the EA0 and EA1 models had insufficient performance for implementation therefore the approach pivoted to create a single EA model EA2 vs not EA2 (i.e., EA0 or EA1)

Figure 3: Performance of Large Language Models on Prospective Note Sets Selected for Implementation

Figure 3a: NYUTron D0 and D2 Models

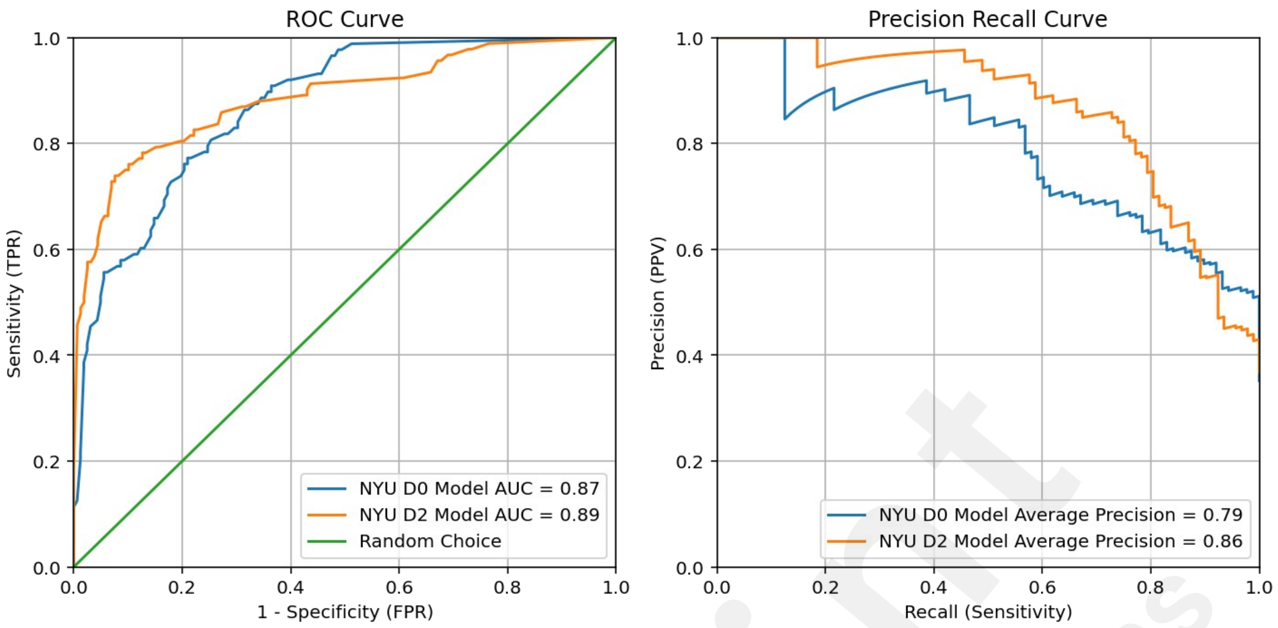


Figure 3b: NYUTron EA2 Binary Model

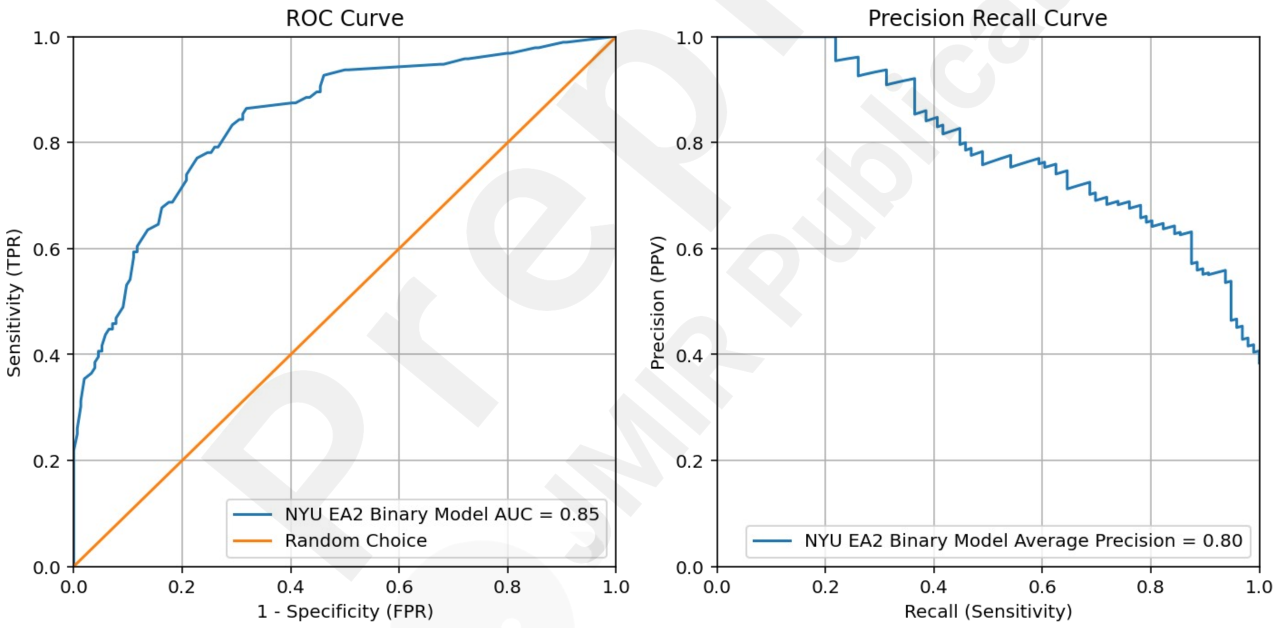
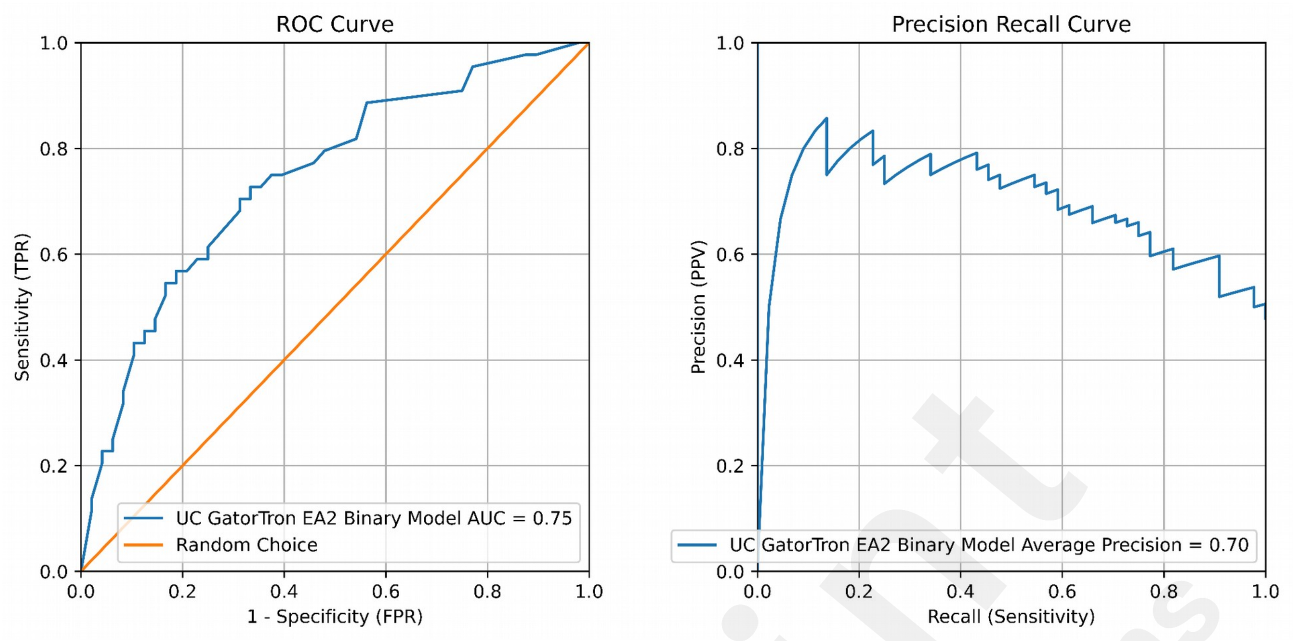


Figure 3c: UC GatorTron EA2 Binary Model


ROC Curve
Sensitivity (TPR)
1 - Specificity (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
UC GatorTron EA2 Binary Model AUC = 0.75
Random Choice
Precision Recall Curve
Precision (PPV)
Recall (Sensitivity)
UC GatorTron EA2 Binary Model Average Precision = 0.70

Declarations

Availability of data and materials: This study involves identified clinical information protected under HIPAA regulations and is not available for distribution. Please email the authors with any inquiries about the data or models.

Disclosures: No authors have conflicts of interest to disclose relevant to this manuscript.

Funding: National Board of Medical Examiners Edward J. Stemmler M.D. Medical Education Research Grant Program

Acknowledgements:

The authors would like to acknowledge Stephen Johnson and Marc Triola for their consultation on methodology. We would also like to thank Lezhi Li for her help with EHR data extraction at the University of Cincinnati College of Medicine.

References

1. National Academies of Sciences, Engineering, Medicine. *Improving Diagnosis in Health Care.* National Academies Press; 2015.
2. Cadieux DC, Goldszmidt M. It's Not Just What You Know: Junior Trainees' Approach to Follow-up and Documentation. *Med Educ.* 2017;51(8):812-825.
3. Singh H, Giardina TD, Meyer AN, Forjuoh SN, Reis MD, Thomas EJ. Types and Origins of Diagnostic Errors in Primary Care Settings. *JAMA Intern Med.* 2013;173(6):418-425.
4. Schiff GD, Bates DW. Can Electronic Clinical Documentation Help Prevent Diagnostic Errors? *N Engl J Med.* 2010;362(12):1066-1069.
5. Baker EA, Ledford CH, Fogg L, Way DP, Park YS. The IDEA Assessment Tool: Assessing the Reporting, Diagnostic Reasoning, and Decision-Making Skills Demonstrated in Medical Students' Hospital Admission Notes. *Teach and Learn in Med.* 2015;27(2):163-173.
6. Schaye V, Miller L, Kudlowitz D, et al. Development of a Clinical Reasoning Documentation Assessment Tool for Resident and Fellow Admission Notes: a Shared Mental Model for Feedback. *J Gen Intern Med.* 2022;37(3):507-512.
7. Brenner AM, Beresin EV, Coverdale JH, et al. Time to Teach: Addressing the Pressure on Faculty Time for Education. *Acad Psych.* 2018;42(1):5-10.
8. Lin SY, Shanafelt TD, Asch SM. Reimagining Clinical Documentation With Artificial Intelligence. *Mayo Clin Proc.* 2018;93(5):563-565.
9. Masters K. Artificial Intelligence in Medical Education. *Med Teach.* 2019;41(9):976-980.
10. Dias RD, Gupta A, Yule SJ. Using Machine Learning to Assess Physician Competence: A Systematic Review. *Acad Med.* 2019;94(3):427-439.
11. Cooper A, Rodman A. AI and Medical Education - A 21st-Century Pandora's Box. *N Engl J Med.* 2023;389(5):385-387.

12. Turner L, Hashimoto DA, Vasisht S, Schaye V. Demystifying AI: Current State and Future Role in Medical Education Assessment. *Acad Med.* 2024 Apr 1;99(4S Suppl 1):S42-S47.

13. Salt J, Harik P, Barone MA. Leveraging Natural Language Processing: Toward Computer-Assisted Scoring of Patient Notes in the USMLE Step 2 Clinical Skills Exam. *Acad Med.* 2019;94(3):314-316.

14. Cianciolo AT, LaVoie N, Parker J. Machine Scoring of Medical Students' Written Clinical Reasoning: Initial Validity Evidence. *Acad Med.* 2021;96(7):1026-1035.

15. Jani KH, Jones KA, Jones GW, Amiel J, Barron B, Elhadad N. Machine Learning to Extract Communication and History-taking Skills in OSCE Transcripts. *Med Educ.* 2020;54(12):1159-1170.

16. Sarker A, Klein AZ, Mee J, Harik P, Gonzalez-Hernandez G. An Interpretable Natural Language Processing System for Written Medical Examination Assessment. *J Biomed Inform.* 2019;98:103268.

17. Schaye V, Guzman B, Burk-Rafel J, et al. Development and Validation of a Machine Learning Model for Automated Assessment of Resident Clinical Reasoning Documentation. *J Gen Intern Med.* 2022;37(9):2230-2238.

18. Feldman J, Hochman KA, Guzman BV, Goodman A, Weisstuch J, Testa P. Scaling Note Quality Assessment Across an Academic Medical Center with AI and GPT-4. *NEJM Catalyst Innov in Care Deliv.* 2024;5(5):CAT. 23.0283.

19. Schaye V, Triola MM. The Generative Artificial Intelligence Revolution: How Hospitalists Can Lead the Transformation of Medical Education. *J Hosp Med.* 2024.

20. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. *Acad Med.* 2024 Jan 1;99(1):22-27.

21. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA.* 2023 Jul 3;330(1):78-80.

22. Cabral S, Restrepo D, Kanjee Z, et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Intern Med.* 2024 May 1;184(5):581-583.

23. Strong E, DiGiammarino A, Weng Y, et al. Chatbot vs Medical Student Performance on Free-response Clinical Reasoning Examinations. *JAMA Intern Med.* 2023;183(9):1028-1030.

24. Restrepo D, Rodman A, Abdulnour RE. Conversations on Reasoning: Large Language Models in Diagnosis. *J Hosp Med.* 2024;19(8):731-735.

25. Goodman KE, Yi PH, Morgan DJ. AI-Generated Clinical Summaries Require More Than Accuracy. *JAMA.* 2024;331(8):637-638.

26. Wu J, Liu X, Li M, et al. Clinical Text Datasets for Medical Artificial Intelligence and Large Language Models—a Systematic Review. *NEJM AI.* 2024;1(6).

27. Hager P, Jungmann F, Holland R, et al. Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-making. *Nat Med.* 2024;30(9):2613-2622.

28. Hswen Y, Abbasi J. AI Will—and Should—Change Medical School, Says Harvard's Dean for Medical Education. *JAMA.* 2023 Nov 21;330(19):1820-1823.

29. NYU Grossman School of Medicine. Institute for Innovations in Medical Education. https://med.nyu.edu/departments-institutes/innovations-medical-education/. Accessed September 17, 2024.

30. NYU Grossman School of Medicine. Predictive Analytics Unit. https://med.nyu.edu/centers-programs/healthcare-innovation-delivery-science/predictive-analytics-unit. Accessed September 28, 2024.

31. NYU Grossman School of Medicine. High Performance Computing Core.https://med.nyu.edu/research/scientific-cores-shared-resources/high-performance-computing-core. Accessed September 17, 2024.

32. Triola MM, Pusic MV. The education data warehouse: a transformative tool for health

education research. *J Grad Med Educ.* 2012;4(1):113-115.

33. University of Cincinnati College of Medicine. Department of Biostatistics, Health Informatics, & Data Sciences. https://med.uc.edu/depart/bhd/about-us/overview. Accessed September 19, 2024.

34. Jiang LY, Liu XC, Nejatian NP, et al. Health System-scale Language Models are All-Purpose Prediction Engines. *Nature.* 2023:1-6.

35. Yang X, Chen A, PourNejatian N, et al. A Large Language Model for Electronic Health Records. *NPJ Digit Med.* 2022;5(1):194.

36. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *arXiv preprint arXiv:190207669.* 2019.

37. Segura-Bedmar I, Martínez P, Herrero-Zazo M. Semeval-2013 task 9: Extraction of Drug-drug Interactions from Biomedical Texts (ddiextraction 2013). Paper presented at: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) 2013.

38. Buda M, Maki A, Mazurowski MA. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Netw.* 2018;106:249-259.

39. Tavakol M, Dennick R. The Foundations of Measurement and Assessment in Medical Education. *Med Teach.* 2017;39(10):1010-1015.

40. National Academies of Medicine. https://dxexscholars.nam.edu/scholars/. Accessed September 14, 2024.

Supplementary Appendix

Supplementary Methods on Note Preprocessing and Truncation

The truncation strategy implemented at NYU was beginning with a start keyword and ending 321 tokens after the start keyword. A list of start keywords such as "presents with", "found to have", or "transferred" (phrases commonly used at the beginning of the assessment and plan) were compiled by human reviewers.[1] The termination after 321 tokens represented the median count of tokens in the truncated sections following the start keyword in our retrospective dataset.

At UC, truncation used the same start keyword approach with one additional start keyword "medical problems being addressed." Similarly to NYU, optimal token length following the start keyword was determined through expert annotation of notes to identify the portion showing reasoning of the primary presenting problem. However, we found that the expert annotated optimal token length varied significantly between notes at UC and tended to be relatively short. We found that a ratio of the original length serves as a more accurate estimate of the optimal length, leading us to adopt a ratio-based approach. The truncation strategy implemented, which involved truncating down to 35% of the token length after the start keyword, winsorized between 86 and 283 tokens, the first quartile, and the maximum optimal length, showed a promising improvement. The mean absolute error was reduced to 67.60, a significant improvement compared to the fixed length of 321, which yielded 191.85.

Supplementary Methods on Type NER Evaluation Metric

The evaluation of system errors can be broken down into five distinct metrics, each representing a different type of error:

1. Correct (COR): This occurs when the system's output and the golden annotation are identical. For instance, if the system identifies "Aspirin" as a drug and the golden annotation also identifies "Aspirin" as a drug, it is considered Correct (COR).
2. Incorrect (INC): This signifies a discrepancy between the system's output and the golden annotation. For example, if the system identifies "Aspirin" as a brand, but the golden annotation identifies "Aspirin" as a drug, it is considered Incorrect (INC).
3. Partial (PAR): This shows that the system's output and the golden annotation have some similarities, but they are not identical. For instance, if the system identifies "Aspirin tablet" as a drug and the golden annotation identifies "Aspirin" as a drug, it is considered Partial (PAR).
4. Missing (MIS): This means that the system failed to capture a golden annotation. For example, if the system fails to identify "Aspirin" as a drug when the golden annotation does, it is considered Missing (MIS).
5. Spurious (SPU): This occurs when the system generates a response that is not present in the golden annotation. For instance, if the system identifies "Paracetamol" as a drug when the golden annotation does not, it is considered Spurious (SPU).

The Type evaluation schema, in particular, is calculated by determining the degree of overlap between the system-tagged entity and the golden annotation. For instance, if the system identifies "Aspirin tablet" as a drug and the golden annotation identifies "Aspirin" as a drug, the Type evaluation schema would consider this a match because there is some overlap between the system-tagged entity and the golden annotation.

To calculate precision, recall, and F1-score for each evaluation schema, two additional quantities need to be calculated:

1. Possible (POS) = COR + INC + PAR + MIS = True Positive (TP) + False Negative (FN)
2. Actual (ACT) = COR + INC + PAR + SPU = True Positive (TP) + False Positive (FP)

Precision is the ratio of correctly identified named-entities by the Named Entity Recognition (NER) system, while recall is the ratio of named-entities in the golden annotations that the NER system correctly retrieved.

1. Precision = (COR + .5 × PAR) / ACT = TP / (TP + FP)
2. Recall = (COR + .5 × PAR)/POS = COR / ACT = TP / (TP + FN)

For example, if the system identifies 10 entities correctly (COR=10), makes 5 incorrect identifications (INC=5), partially identifies 3 entities (PAR=3), misses 2 entities (MIS=2), and identifies 4 spurious entities (SPU=4), the Possible (POS) and Actual (ACT) quantities would be calculated as follows:

POS = COR + INC + PAR + MIS = 10 + 5 + 3 + 2 = 20
ACT = COR + INC + PAR + SPU = 10 + 5 + 3 + 4 = 22

Then, the precision and recall for exact and partial matches would be calculated as follows:

Precision = (COR + .5 × PAR) / ACT = (10 + .5 * 3) / 22 = .48
Recall = (COR + .5 × PAR)/POS = (10 + .5 * 3) / 20 = .52

Supplementary Methods on Model Hyperparameters and Model Packages

At NYU, each LLM was fine-tuned using the following hyperparameters: a training epoch number of 50, a training batch size of 64, a learning rate of 0.0001, weight decay of 0.01, and a dropout of 0.1. Python 3.8.5 was used for all of our data extraction, model development and validation. Several open-source libraries were utilized: spaCy 3.4.1, scispaCy 0.5.1, scikit-learn 1.0.2, nltk 3.6.5, pandas 1.4.2, HuggingFaceTransformers 4.25.1, torch 1.10.0, and tensorflow 2.5.3.

At UC, the hyperparameters were as follows: a training epoch number of 50, a training batch size of 16, a learning rate of 0.000005 for the encoder layers, 0.0005 for the classifier layer, a weight decay of 0.05, and a dropout of 0.1. Additionally, learning rates were adjusted using an exponential warm-up of 10 epochs and exponential decrease with a 0.9 gamma afterward. Python 3.8.10, scikit-learn 1.3.2, nltk 3.8.1, pandas 1.5.1, HuggingFaceTransformers 4.34.1, torch 2.1.0 were used.

Supplementary Results on Selecting Thresholds

When selecting the optimal thresholds for all LLM models, both Positive Predictive Value (PPV) and Negative Predictive Value (NPV) that were greater than 69% and False Positives (FP) that are approximately equal to False Negatives (FN) were the major considerations. Such an approach to find the optimal threshold yields balanced errors, which could potentially enhance resident acceptability without compromising the model's performance. An additional advantage of this threshold is its calibration, as the positive and negative rates would closely mirror the actual prevalence of D and EA scores. This implies that the scores aggregated at the trainee level across multiple measures would likely reflect the ground truth accurately. On the NYUTron models the following probability cutoffs were selected for new, unseen data: 0.39 for D0 model, 0.23 for D2 model, and 0.54 for EA2 model, and on the UC EA2 GatorTron model 0.89.

References

1. Schaye V, Guzman B, Burk-Rafel J, et al. Development and Validation of a Machine Learning Model for Automated Assessment of Resident Clinical Reasoning Documentation. *J Gen Intern Med.* 2022;37(9):2230-2238.