PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Evaluating Large Language Models in extracting cognitive exam dates and scores

Hao Zhang[1], Neil Jethani[1], Simon Jones[1], Nicholas Genes[1], Vincent J. Major[1], Ian S. Jaffe[1], Anthony B. Cardillo[1], Noah Heilenbach[1], Nadia Fazal Ali[1], Luke J. Bonanni[1], Andrew J. Clayburn[1], Zain Khera[1], Erica C. Sadler[1], Jaideep Prasad[1], Jamie Schlacter[1], Kevin Liu[1], Benjamin Silva[1], Sophie Montgomery[1], Eric J. Kim[1], Jacob Lester[1], Theodore M. Hill[1], Alba Avoricani[1], Ethan Chervonski[1], James Davydov[1], William Small[1], Eesha Chakravartty[1], Himanshu Grover[1], John A. Dodson[1], Abraham A. Brody[1,2], Yindalon Aphinyanaphongs[1], Arjun Masurkar[1], Narges Razavian[1]*

1 NYU Grossman School of Medicine, New York, New York, United States of America, **2** Rory Meyers College of Nursing, New York University, New York, New York, United States of America

* narges.razavian@nyulangone.org

Citation: Zhang H, Jethani N, Jones S, Genes N, Major VJ, Jaffe IS, et al. (2024) Evaluating Large Language Models in extracting cognitive exam dates and scores. PLOS Digit Health 3(12): e0000685. https://doi.org/10.1371/journal.pdig.0000685

Editor: Imon Banerjee, Mayo Clinic, Arizona, UNITED STATES OF AMERICA

Received: March 1, 2024

Accepted: October 28, 2024

Published: December 11, 2024

Data Availability Statement: The clinical notes used for this study were collected from the NYU Langone Health System EHR maintained by the NYULH Datacore team. These clinical notes contain potentially identifying or sensitive patient information, and according to the Institutional Review Board and Data Sharing Committee at NYU Langone, cannot be made publicly available. Researchers interested in the data used in this study should submit a reasonable request to the data sharing committee datasharing@nyulangone.org, and the request will undergo institutional

Abstract

Ensuring reliability of Large Language Models (LLMs) in clinical tasks is crucial. Our study assesses two state-of-the-art LLMs (ChatGPT and LlaMA-2) for extracting clinical information, focusing on cognitive tests like MMSE and CDR. Our data consisted of 135,307 clinical notes (Jan 12th, 2010 to May 24th, 2023) mentioning MMSE, CDR, or MoCA. After applying inclusion criteria 34,465 notes remained, of which 765 underwent ChatGPT (GPT-4) and LlaMA-2, and 22 experts reviewed the responses. ChatGPT successfully extracted MMSE and CDR instances with dates from 742 notes. We used 20 notes for fine-tuning and training the reviewers. The remaining 722 were assigned to reviewers, with 309 each assigned to two reviewers simultaneously. Inter-rater-agreement (Fleiss' Kappa), precision, recall, true/false negative rates, and accuracy were calculated. Our study follows TRIPOD reporting guidelines for model validation. For MMSE information extraction, ChatGPT (vs. LlaMA-2) achieved accuracy of 83% (vs. 66.4%), sensitivity of 89.7% (vs. 69.9%), true-negative rates of 96% (vs 60.0%), and precision of 82.7% (vs 62.2%). For CDR the results were lower overall, with accuracy of 87.1% (vs. 74.5%), sensitivity of 84.3% (vs. 39.7%), true-negative rates of 99.8% (98.4%), and precision of 48.3% (vs. 16.1%). We qualitatively evaluated the MMSE errors of ChatGPT and LlaMA-2 on double-reviewed notes. LlaMA-2 errors included 27 cases of total hallucination, 19 cases of reporting other scores instead of MMSE, 25 missed scores, and 23 cases of reporting only the wrong date. In comparison, ChatGPT's errors included only 3 cases of total hallucination, 17 cases of wrong test reported instead of MMSE, and 19 cases of reporting a wrong date. In this diagnostic/prognostic study of ChatGPT and LlaMA-2 for extracting cognitive exam dates and scores from clinical notes, ChatGPT exhibited high accuracy, with better performance compared to LlaMA-2. The use of LLMs could benefit dementia research and clinical care, by identifying eligible patients for treatments initialization or clinical trial enrollments. Rigorous evaluation of LLMs is crucial to understanding their capabilities and limitations.

review and will be subject to local and national ethical approvals.

Funding: This study was supported by NYU Langone Medical Center Information Technology (MCIT) center, National Institute On Aging, of the National Institutes of Health (R01AG085617 to NR and AM, P30AG066512 to NR and AM, P30AG066512 to HZ, SJ, VJM, JAD, AAB, YA, AM, and NR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Large-scale language models (LLMs) have emerged as powerful tools in natural language processing (NLP), capable of performing diverse tasks when prompted. Since reliable performance of LLMs in clinical tasks is essential, our study evaluates two advanced LLMs—ChatGPT and LlaMA-2—for their ability to extract clinical information from health records, particularly cognitive test results: MMSE (Mini-Mental State Examination) and CDR (Clinical Dementia Rating). We analyzed 765 clinical notes from over a decade, focusing on how well these models could identify and date specific test mentions. ChatGPT accurately extracted MMSE and CDR details from most notes, demonstrating greater accuracy, sensitivity. Our findings also reveal common errors, with ChatGPT generally producing fewer inaccuracies. This research emphasizes the potential for LLMs to enhance dementia research and patient care by improving the identification of eligible patients for treatment or clinical trials. However, a thorough understanding of these models' strengths and weaknesses is essential for their effective application in real-world clinical settings.

Introduction

Large-scale language models (LLMs) [1–4] have emerged as powerful tools in natural language processing (NLP), capable of performing diverse tasks when prompted [5,6]. These models have demonstrated impressive clinical reasoning abilities [7], successfully passing medical licensing exams [8–10] and generating medical advice on distinct subjects, including cardiovascular disease [11], breast cancer [12], colonoscopy [13], and general health inquiries [6,14–16]. These models can produce clinical notes [16] and assist in writing research articles [16]. Medical journals have begun developing policies around use of LLMs in writing [17–22] and reviewing. Examples of such LLMs include ChatGPT [1,2], Med-PALM-2 [3], LlaMA-2 [4], and open-source models actively produced by the community [23].

In this study, we focus on evaluating *information extraction* abilities of Large Language Models from clinical notes, specifically focusing on proprietary ChatGPT (powered by GPT-4 [2]), and open source LlaMA-2 [4] LLMs. Information extraction involves the retrieval of specific bits of information from unstructured clinical notes, a task historically handled by rule-based systems [24–30] or language models explicitly trained on datasets annotated by human experts [31–36]. Rule-based systems lack a contextual understanding and struggle with complex sentence structures, ambiguous language, and long-distance dependencies, often leading to high false positive rates and low sensitivities [37–40]. Additionally, training a new model for this task can be computationally demanding and require substantial human effort. In contrast, LLMs, such as ChatGPT or LlaMA-2, operate at "zero-shot" capacity [41–43], i.e., only requiring a prompt describing the desired information to be extracted.

Despite their promise, LLMs also have a potential limitation—the generation of factually incorrect yet highly convincing outputs, commonly known as "hallucination." The massive architectures and complex training schemes of LLMs hamper "model explanation" and the ability to intrinsically guarantee behavior. This issue has been extensively discussed in the literature, emphasizing the need for cautious interpretation and validation of information generated by LLMs [2,44,45].

One area where LLMs may greatly benefit healthcare is in the identification of memory problems and other symptoms indicative of Alzheimer's Disease and Alzheimer's Disease

Related Dementias (AD/ADRD) within clinical notes. AD/ADRD is commonly underdiagnosed or diagnosed later in the disease trajectory, particularly in racial and ethnic minoritized groups [46–51]. The precise extraction of cognitive test scores holds significant importance in the development and clinical validation of tools that can facilitate early detection [52] of AD/ADRD in the clinic. Earlier identification can lead to a host of benefits, including assisting with advanced care planning, performing secondary cardiovascular disease prevention, which may reduce worsening of cognitive impairment [53,54], identification for serving in research trials [55–57], and with the rapid advancement in biologic therapeutics, the opportunity to receive potentially disease modifying drugs [57,58]. Accurately extracting cognitive exam scores (often buried in clinical notes and not documented in any structured field), enables validation, training and fine-tuning of models at a much larger scale in a clinical setting for a much more racial/ethnically diverse patient population set compared to current research cohorts.

The primary focus of this paper is therefore on the validation of two state-of-the-art LLMs (ChatGPT powered by GPT-4, and LlaMA-2), for information extraction related to cognitive tests, specifically the Mini-Mental State Examination (MMSE) [59] and Clinical Dementia Rating (CDR) [60], from clinical notes of a racially and ethnically diverse patient population. Our objective is to accurately extract all instances of (the exam score, and the date when the exam was administered) using these LLMs.

This study represents a large-scale formal evaluation of two state of the art LLMs (ChatGPT, and LlaMA-2) performance in information extraction from clinical notes. Going forward, we intend to employ this benchmark dataset to validate other (open or closed-source) LLMs. Furthermore, we plan to adopt a similar approach to validate LLMs for information extraction across various clinical use cases. By prioritizing prompt engineering with ChatGPT and LlaMA-2 for extracting clinical information, this research aims to enhance our understanding of the potential of LLMs in healthcare and facilitate the development of reliable and robust clinical information extraction tools.

Methods

This study is approved under IRB i20-01095, "Understanding and predicting Alzheimer's Disease." NYU DataCore services were utilized to prepare the data as described below. A HIPAA-compliant private instance of ChatGPT (Microsoft Azure OpenAI Service) was utilized for this study. LlaMA-2 ("Llama-2-70b-chat" version) was evaluated on two A100 Nvidia GPUs on our local high performance computing servers. This Diagnostic/Prognostic study designed to validate the diagnostic accuracy of two LLMs (ChatGPT and LlaMA-2) in extracting cognitive exam dates and scores, follows the TRIPOD Prediction Model Validation reporting guidelines (S1 Checklist) [61].

Dataset

An original cohort of 135,307 clinical notes corresponding to inpatient, outpatient, and emergency department visits between January 12th 2010 and May 24th 2023, which included any of the following keywords ('MMSE', 'CDR,' or 'MoCA' case-insensitive) were identified (see Fig 1). MMSE stands for Mini Mental State Exam, CDR stands for Cognitive Dementia Rating, and MoCA stands for Montreal Cognitive Assessment [62]. These notes belonged to 52,948 patients. From among these patients, 26,355 had a non-contrast brain Magnetic Resonance Imaging (MRI) in the system. Limiting the clinical notes to those who had an MRI in the system resulted in 77,547 notes. After extracting the notes, we further excluded 43,082 notes that only mentioned MoCA, yielding 34,465 clinical notes for analysis.

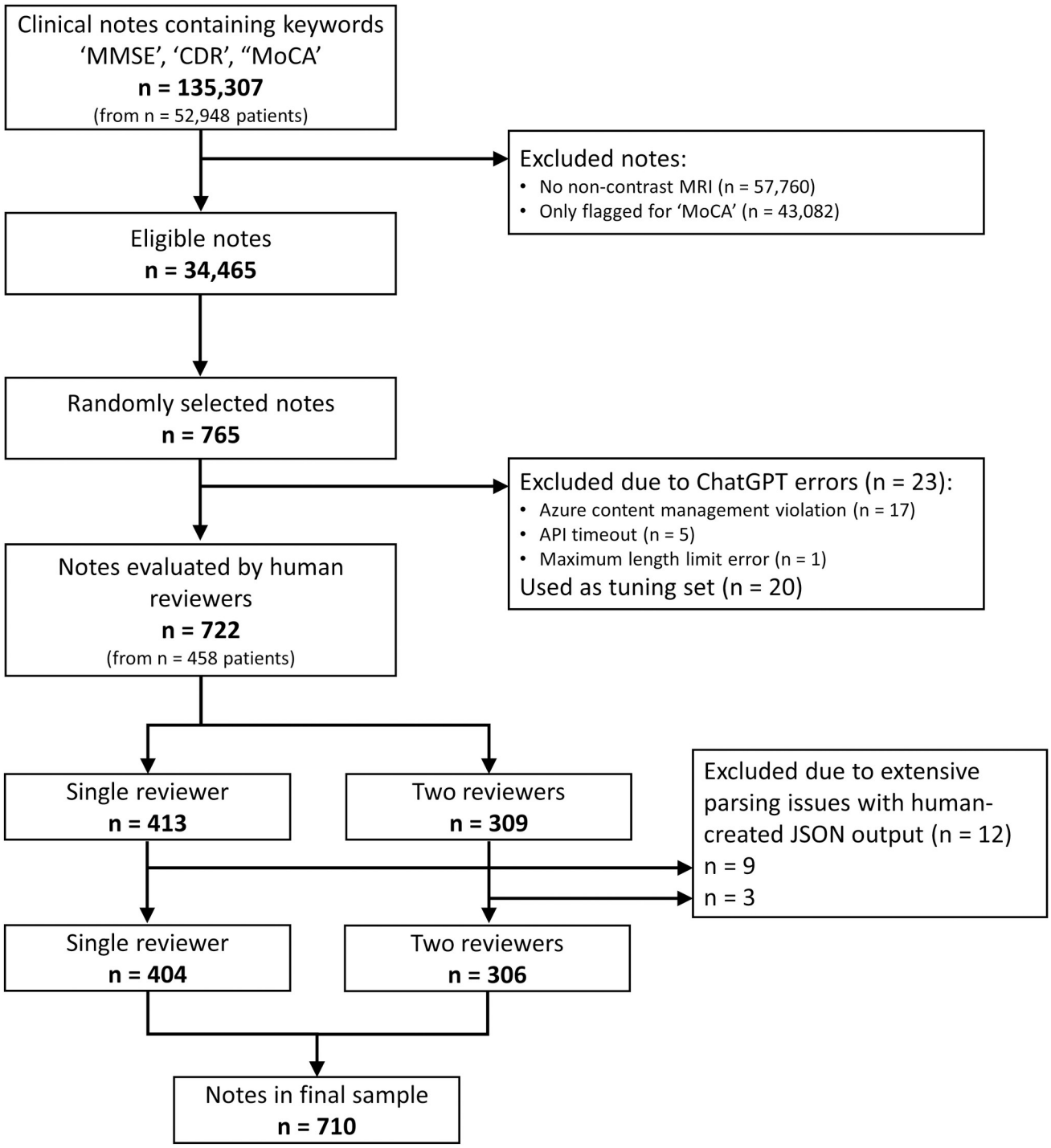

Fig 1. Flowchart of clinical notes evaluated for inclusion in the final sample of GPT-analyzed notes.

https://doi.org/10.1371/journal.pdig.0000685.g001

The choice for requiring patients to have a brain MRI as well as MMSE and/or CDR enables us to have a similar level of granularity as the Alzheimer's Disease Neuro-Imaging Initiative (ADNI) [63], which also uses MMSE and CDR for definition of mild cognitive impairment and dementia stages. This further enables us to harmonize our clinical dataset with these large

research cohorts. To elucidate the impact of this choice (restriction of cohort to those with MRI) on the racial breakdown of our study, we include a demographics comparison between the two sets (original 52,948 patients, and the 26,355 with an MRI) in S1 Section. Similarly, the choice to ignore MoCA was due to the lack of inclusion of MoCA in standard definition for stages of cognitive impairment in ADNI. The mild cognitive impairment and (mild, moderate or severe) dementia definition criteria utilized in ADNI are included in S1 Table. Data harmonization is beyond the scope of this paper, although information extraction plays a substantial role in enabling it.

From among 34,465 notes that fit the inclusion criteria, a random selection of 765 notes was identified to undergo information extraction via ChatGPT and manual evaluation. 765 was the total number of the notes needed to satisfy two conditions: 1) Each reviewer not being assigned more than 50 notes to review, and 2) at least around 15 notes per reviewer being double-reviewed by another random reviewer. From among these 765 notes, ChatGPT encountered application programming interface (API) errors in 23 cases (3%). These errors arose from "Azure content management violations" [64] (17 cases), API timeouts (5 cases), and maximum length limit errors (1 case). S2 Table includes a more detailed description of these errors. The remaining 742 were considered for assignment to domain expert reviewers, and underwent analysis by LlaMA-2.

Generative AI, ChatGPT

A private, HIPAA-compliant instance of ChatGPT (GPT-4, API version "2023-03-15-preview") was used on these 765 notes to extract all instances of the cognitive tests—MMSE and CDR—along with the dates at which the tests were mentioned to have been administered. Examples of our task are provided in the S2 Section. Inference was successful for 742 notes. The complete API call, along with the exact prompt, the temperature, and other hyper-parameters are included in S3 Table. The prompt included a request to return these results in a JSON format. ChatGPT's response (full), as well as the JSON formatted dialogue response were recorded in one session on June 9th 2023. The notes sent to ChatGPT were text-only, stripped of the rich-text formatting (RTF) native to our EHR system (Epic Systems, Verona, WI). This reduced token count by approximately ten-fold, enabling notes to fit into the GPT4-8K input window and removing a substantial source of confusion for the LLM in prompt tuning. The date that the encounter was recorded in Epic was appended at the beginning of the note, proceeding with a column (":") then the note text. See S3 Table for the API request, including the prompt.

Generative AI, LlaMA-2

We used LlaMA-2 (version "Llama-2-70b-chat") on all the notes which ChatGPT produced valid answer. All pre-processing steps on the notes were similar to that of ChatGPT. The context window was limited to the first 3696 tokens. The complete API call, along with the exact prompt, the temperature, and other hyper-parameters are included in S4 Table.

Hyper-parameter and prompt tuning

For both ChatGPT and LlaMA-2, we assigned 20 notes out of the 742 as our hyper-parameter and prompt tuning set. For ChatGPT, an interactive cloud-based environment (i.e playground) was utilized initially to fine-tune the prompt. After initial exploratory analysis using these 20 notes, they were scored via the API using the best prompt and hyper-parameter found in the interactive mode. For LlaMA-2, the exploration was performed locally, on the same 20 notes. For both models, we explored the following model parameters: max_token_length,

temperature. All human expert reviewers (detailed below) were instructed to first review the ChatGPT results of the 20 cases in a RedCap survey. The goal of this step was to train the reviewers, refine the information presented in RedCap, improve clarification of the questions, and potentially refine the prompt. These 20 notes were then excluded from any additional analysis.

Human expert reviewers

Our team included 22 medically trained expert reviewers who volunteered and were trained to review an (HTML formatted) note, provide ground truth, and judge the correctness and completeness of ChatGPT answers for each cognitive test. Fully (HTML) formatted notes were pulled using an Epic web service, and were fed into the RedCap survey. Redcap survey rendered the note's HTML formatting, to ensure notes could be displayed to users in the same format as the readers are accustomed to seeing them clinically, rather than the text-only, computer-friendly format provided to GPT. To generate ground-truth, the reviewers used ChatGPT responses as the basis, and corrected any errors ChatGPT made.

For 21 of these reviewers, each reviewer was assigned approximately 50 clinical notes to evaluate. From among each reviewer's 50 assigned notes, about 15 notes were assigned to another random reviewer. The assignment algorithm randomly selected a pair of reviewers for each of our 309 double-reviewed notes and assigned the remaining notes to a randomly selected reviewer until each reviewer reached 50 notes or we fully assigned all notes. This random assignment was a necessary step for ensuring correctness of Fleiss' Kappa [65] metric for inter-rater-agreement. As a result, there was a slight variation in the total number of assigned notes for each reviewer.

Overall, 722 notes were assigned to these 21 reviewers, of which 309 were double-reviewed and 413 were solo-reviewed. The double-reviewed 309 notes were utilized in reporting inter-rater-agreement metrics. After the review, 69 out of 309 notes had at least one disagreement between the two reviewers based on one of the four questions: *Whether ChatGPT's response on MMSE was correct; whether ChatGPT's response on MMSE included all instances of MMSE found in the clinical note; whether ChatGPT's response on CDR was correct; and whether ChatGPT's response on CDR included all instances of CDR found in the clinical note*. A 22nd reviewer was then tasked to review these 69 notes again to provide a third review. Majority vote was then employed to identify the final answer and the ground truth provided by the reviewer whose answer was in the majority vote was used to calculate detailed precision/recall metrics. When both reviewers fully agreed and their JSON results were both valid for analysis, we randomly selected one to compute the precision and recall. Details of the parsing of the JSON result are included in the S3 Section. These expert-provided ground truth results were the basis for evaluating LlaMA-2.

Statistical approach

We reported Fleiss' Kappa [65] as a measure of inter-rater-agreement for double-reviewed notes. We reported this metric for the four questions on ChatGPT-generated responses (Is MMSE complete/correct, and is CDR complete/correct). Additionally, for double-reviewed notes, we derived inter-rather-agreement by computing 2-way Fleiss' Kappa for MMSE and CDR lists of (outcome and date) tuples extracted from the JSON responses by expert reviewers. Fleiss' Kappa is useful when the assignment of a note to reviewer pairs has been random (uniform), and each note has been reviewed by a subset of reviewers [66,67]. Only exact outcome and date tuple matches were considered to be in agreement between raters (i.e [MMSE-27/30, date "10-10-2010"], with [MMSE-26/30, date "10-10-2010"] is just as bad as [MMSE-5/30, date

"10-10-2012"]). We also report a 3-way Fleiss' Kappa on the entries of MMSE and CDR results extracted from the JSON results, computing the joint agreement between the results of ChatGPT and the results provided by two human reviewers.

We also report per test type (MMSE and CDR), Accuracy, True and False Negative Rates, Micro- and Macro-Precision and Micro- and Macro-Recall for both ChatGPT and LlaMA-2. Accuracy is defined as the percentage of correct results (at clinical note level), correct being defined as the list of (Value/Date) tuples in the JSON entries for the LLM and Ground Truth being fully identical. Macro-Precision for MMSE (or CDR) is the average (at the note level) of percentage of correct MMSE (or CDR) tuples extracted (correct both in date and score values compared to an entry mentioned in the ground truth for MMSE (or CDR)). Macro-Recall for MMSE (or CDR) is the average (at the note level) of the percentage of the MMSE items in the ground truth that are extracted by the LLM. *Micro*-precision is calculated as percentage of *correct* MMSE (or CDR) items extracted by the LLM, from among all extracted MMSE (or CDR) items by that LLM, and is calculated as one number across all notes combining all notes' entries. Micro-recall is similarly calculated as the percentage of all MMSE (or CDR) items mentioned in the ground truth that were extracted by the LLM.

Results

ChatGPT analyzed 765 notes for extraction of Mini Mental Status Exam (MMSE) and Cognitive Dementia Rating (CDR) scores and exam dates. Of these, 23 encountered API error (3%), and 20 were used to fine-tune prompt and hyper-parameters. The remaining 722 notes were assigned to human expert reviewers who manually reviewed (and provided ground truth for) these notes. LlaMA-2 analyzed these 722 notes as well. Characteristics of these 722 notes and associated patients are included in Table 1.

Of the double-reviewed 309 notes, 69 had at least one disagreement between the responses to the four questions (if ChatGPT's response for MMSE/CDR is correct/complete) and were assigned to a new reviewer for a third opinion. Among the responses with disagreement, 9 disagreed about correctness of MMSE answers, 40 disagreed about completeness of MMSE answers, 17 disagreed about correctness of CDR answers, and 22 disagreed about completeness of CDR answers. The average response (at the note level) by the included reviews for the four yes/no questions are included in Table 2. Overall reviewers considered ChatGPT's response to be 96.5% and 98% correct for MMSE and CDR respectively. The assessment for whether ChatGPT's answers are also complete (i.e. they do not miss anything) was slightly lower averaging about 84% and 83% for MMSE and CDR respectively.

The inter-rater-agreements between reviewers were calculated based on Fleiss' Kappa and are summarized in Table 3. In addition to measuring Fleiss' Kappa between reviewers based on double-reviewed notes (reported as 2-way Fleiss' Kappa in Table 3), we also report agreement between ChatGPT, and the two human reviewers (reported as 3-way Fleiss' Kappa in Table 3). The 2-way agreement on the yes/no questions was high (94% agreement between reviewers for MMSE and 89% agreement for CDR). There was some disagreement in judging the completeness of the answer, leading to a Kappa value of 75% for MMSE (and 85% for CDR). More notably, when analyzing the elements of the ground truth JSON, the 2-way agreement was excellent both for scores (83% for MMSE and 80% for CDR) and for dates (93% for MMSE and 79% for CDR). When measuring the 3-way agreement, there was an increase in all the metrics except MMSE dates. The accuracy and results of JSON formatting of the responses are included in S4 Section.

ChatGPT had an excellent True Negative Rate—over 96% for MMSE and 100% for CDR in double-reviewed notes (**Table 4**). Both results had high recall (sensitivity), reaching 89.7% for

Table 1. Characteristics of 722 notes which are manually evaluated, and their corresponding patients.

Feature	All notes (N = 722 notes from 458 patients)	Double reviewed notes (N = 309 notes from 236 patients)
Patient demographics		
Age at time of note (mean (sd))	72.64 (14.01)	73.68 (14.01)
Gender		
Female (%)	242 (52.84%)	124 (52.54%)
Male(%)	216 (47.16%)	112 (47.46%)
Race		
Asian	27 (5.90%)	10 (4.24%)
Black	39 (8.52%)	17 (7.20%)
White	334 (72.93%)	178 (75.42%)
American Indian	1 (0.22%)	0 (0.00%)
Unknown	57 (12.45%)	31 (13.14%)
Note characteristics		
Date ranges (min to max)	2011/11/21 to 2023/05/10	2011/11/21 to 2023/05/10
Length (in words) (mean (SD))	8428.2 (3822.3)	8306.2 (3851.1)
ChatGPT (Prompt Tokens)	2212.93 (1002.9)	2174.9 (992.3)
ChatGPT (Completion Tokens)	64.3 (49.6)	64.2 (46.5)
ChatGPT (Total Tokens)	2277.3 (1017.9)	2239.1 (1005.0)
Llama2 (Prompt Tokens)	2860.8 (1224.2)	2810.4 (1208.4)
Llama2 (Completion Tokens)	140.2 (112.8)	146.9 (125.3)
Llama2 (Total Tokens)	3000.9 (1276.7)	2957.4 (1270.8)

https://doi.org/10.1371/journal.pdig.0000685.t001

MMSE (macro-recall) and 91.3% for CDR (macro-recall). MMSE was more frequently mentioned in the notes and ChatGPT's macro precision (PPV) was 82.7%. CDR, on the other hand, was less frequent, and we observed that ChatGPT hallucinates (factitiously generates) results occasionally leading to a macro precision of only 57.5%. LlaMA-2 results were significantly lower than that of ChatGPT across all metrics. A detailed qualitative analysis of the ChatGPT errors for both CDR and MMSE, and LlaMA-2 results for MMSE are included in S5 Section. The majority of the errors corresponded to ChatGPT presenting results of another test instead of the one indicated as the answer. LlaMA-2 had higher rate of unexplained hallucinations. Taking positive and negative results into account, overall, ChatGPT had the highest performance with MMSE and CDR results being 83% and 89% accurate according to the double-reviewed notes.

Table 2. Average response (at the note level) of the responses of reviewers in judging if ChatGPT's answers for MMSE and CDR are correct and/or complete.

	All notes (N = 722)	Double reviewed notes (N = 309)
Is ChatGPT's answer for MMSE correct? (%)	96.5 (sd 18.2)	96.4 (sd 18.5)
Is ChatGPT's answer for MMSE complete? (%)	85.0 (sd 35.7)	84.7 (sd 36.0)
Is ChatGPT's answer for CDR correct? (%)	98.0 (sd 13.7)	99.6 (sd 5.6)
Is ChatGPT's answer for CDR complete? (%)	80.4 (sd 39.6)	83.4 (sd 37.1)

https://doi.org/10.1371/journal.pdig.0000685.t002

Table 3. Fleiss' kappa inter-rater-agreement metric between reviewers (2-way) and reviewers and ChatGPT (3-way) over the double-reviewed notes.

	2-way Fleiss' kappa (Among human reviewers) On N = 309 double-reviewed notes, n = 21 reviewers (%)	3-way Fleiss' Kappa (between ChatGPT and two human reviewers) On N = 309 double-reviewed notes, n = 21 reviewers (%)
Binary Questions		
Is MMSE list generated by ChatGPT correct?	94.2	NA
Is MMSE list generated by ChatGPT complete?	75.2	NA
Is CDR list generated by ChatGPT correct?	89.0	NA
Is CDR list generated by ChatGPT complete?	85.8	NA
Individual (value/date) tuples from ChatGPT and Ground-Truth JSON results.		
MMSE values (of the scores in the note)	83.6	93.7
MMSE dates (of the scores in the note)	93.3	87.2
CDR values (of the scores in the note)	80.5	87.0
CDR dates (of the scores in the note)	79.0	82.5

https://doi.org/10.1371/journal.pdig.0000685.t003

Discussion

In this study, our primary objective was to evaluate the performance of two state of the art LLMs (ChatGPT and LlaMA-2), in extracting information from clinical notes, specifically focusing on cognitive tests such as the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR). Our results revealed that ChatGPT achieves high accuracy in extracting relevant information for MMSE and CDR scores, as well as their associated dates, with high recall, capturing nearly all of the pertinent details present in the clinical notes. The overall accuracy of ChatGPT in information extraction for MMSE and CDR were 83% and 89% respectively. The extraction was highly and had outstanding true-negative-rates. The precision of the extracted information was also high for MMSE although in the case of CDR, we observed that ChatGPT occasionally mistook other tests for CDR. Based on the ground-truth provided by our reviewers, 89.1% of the notes included an MMSE documentation instance, whereas only 14.3% of the notes included a CDR documentation instance. This, combined with our analysis of the errors, explain lower precision in the CDR case, and suggest combining ChatGPT with basic NLP preprocessing may improve the LLM performance further. Compared to ChatGPT, the open-source state of the art LLM (LlaMA-2) achieved lower performance across all metrics. The substantial inter-rater-agreement among our expert reviewers further supported the robustness and validity of our findings, and the reviewers considered ChatGPT's responses correct and complete.

The findings of our study demonstrate that ChatGPT (powered by GPT-4), offer a promising solution for extracting valuable clinical information from unstructured notes. This approach provides a more efficient and scalable approach compared to previous methods that either rely on rigid rule-based systems or involve training resource intensive task specific models. Validated and accurate LLMs such as ChatGPT can be effortlessly applied to enhance the value of clinical data for research, enable harmonization with disease registries

Table 4. Aggregate Accuracy, True Negative Rate, (Micro- and Macro-) Precision and Recall for MMSE and CDR scores extracted by ChatGPT and LlaMA-2.

	All notes with parsed JSON (N = 710)		Double-reviewed notes with parsed JSON (N = 306)	
	ChatGPT	LlaMA-2	ChatGPT	LlaMA-2
MMSE				
Total notes without any MMSE (in ground truth)	115		48	
Total notes without any MMSE (in GPT results)	77	110	25	46
Total correctly predicted empty MMSEs	76	66	24	23
MMSE True Negative Rate (%)	98.7	60.0	96	50.0
MMSE False Negative Rate(%)	1.2	40.0	4	50.0
Remaining notes with un-empty GPT response undergone Precision/Recall calculation for MMSE	633	600	281	260
Total MMSE instances predicted	831	957	366	410
MMSE Macro Precision (mean % (sd %))	82.9 (sd 36.2)	62.2(sd 45.5)	82.7 (sd 36.8)	63.4 (sd 44.9)
MMSE Macro Recall (mean % (sd %))	87.8 (sd 30.4)	69.9 (sd 43.5)	89.7 (sd 28.3)	71.8 (sd 42.1)
MMSE Micro Precision (%)	83.8	57.7	84.1	59.3
MMSE Micro Recall (%)	83.7	68.1	87.5	69.0
Total notes with any error MMSE result	121	238	52	98
Overall accuracy of MMSE (%)	82.9	66.4	83.0	68.0
CDR				
Total notes without CDR (in ground truth)	608		260	
Total notes without CDR (in GPT results)	533	497	233	215
Total correctly predicted empty CDR	532	489	233	212
CDR True Negative Rate (%)	99.8	98.4	100	98.6
CDR False Negative Rate (%)	0.2	1.6	0	1.4
Remaining notes with un-empty GPT response undergone Precision/Recall calculation for CDR	177	213	73	153
Total CDR instances predicted	256	344	92	153
CDR Macro Precision (mean % sd %)	48.3 (sd 49.9)	16.1 (sd 35.5)	57.5 (sd 49.4)	18.1 (sd 36.9)
CDR Macro Recall (mean % sd %)	84.3 (sd 36.3)	39.7 (sd 48.7)	91.3 (sd 28.1)	43.5 (sd 49.6)
CDR Micro Precision (%)	36.3	12.0	51.0	13.2
CDR Micro Recall (%)	85.3	37.6	92.1	39.2
Total notes with any error CDR result	91	181	31	76
Overall accuracy of CDR (%)	87.1	74.5	89.8	75.4

https://doi.org/10.1371/journal.pdig.0000685.t004

and biobanks, improve outreach programs within health centers, and contribute to the advancement of precision medicine. Additionally, the availability of large labeled datasets resulting from this information extraction process can also enable AI models to be trained for a wide variety of tasks.

Furthermore, our findings have implications for future AD/ADRD research. Currently, the majority of research in scalable development and validation of AI tools for early AD/ADRD detection rely on research cohorts. These cohorts are overwhelmingly white (NACC cohort is 83% white [68] ADNI cohort is 92% white [63], and do not represent true at-risk populations who tend to have higher comorbid disease burden [50]. Due to late detection and diagnosis of AD/ADRD [46–49], clinical data often lacked the details necessary for accurate case identification (i.e. structured data such as ICD codes would yield low sensitivities). Using LLMs to extract data from clinical notes has the potential to improve the quality of clinical data, paving the way for clinical validation and development of clinically applicable novel AI tools and performing cognitive-health precision medicine at scale.

Limitations

Our focus was on evaluating the information extraction capabilities of two current state of the art LLMs, specifically ChatGPT powered by GPT-4, and LlaMA-2, rather than comparing it to all other LLMs or NLP methods. We believe that our results may be enhanced with better prompt engineering and combining LLMs with standard NLP. In the future, we hope to include other LLMs to evaluate this task. One limitation during the labeling stage is that we generated ChatGPT responses first before the experts review and correct to create the ground truth. While this could introduce potential bias towards ChatGPT, we believed that ground truth is still valid to evaluate other models. ChatGPT might also not be reliable 100% of the time, as we have seen that it failed to generate responses for a small fraction of the notes. In production, it is critical to have back-up plans in place such as alternative LLMs to ensure the system can reliably extract scores for all notes. Additionally, we conducted a large-scale human evaluation for a single dementia use case, prioritizing result reliability over assessing various clinical scenarios. It is also important to note that our findings pertain specifically to information retrieval from clinical notes and do not predict how LLMs will perform on medical tasks requiring diagnosis, treatment recommendation, or summarization. For the scope of the study, we focused on patients with an MRI exam, and we have seen that there is a distribution difference in patients with an MRI than those tho do not (S1 Table). There might be a potential difference on how clinicians document cognitive scores in the two populations. In future studies, we would like to explore how this difference could affect model performance. Finally, these large language model requires extensive hardware resources, meaning carbon footprint is larger compared to traditional NLP methods. However, as the scores were often discussed in natural language, where the information (type of test, date of test, and test scores) can be far apart from each other, traditional NLP methods are not viable for this particular task without extensive efforts and large amount of training samples. We have included a few examples of clinical texts in S2 Table to demonstrate the heterogeneity of the texts.

Conclusions

In this diagnostic/prognostic study of ChatGPT and LlaMA-2 for extracting cognitive exam dates and scores from clinical notes, ChatGPT exhibited high accuracy in extracting MMSE scores and dates, with better performance compared to LlaMA-2. The use of LLMs could benefit dementia research and clinical care, by identifying eligible patients for treatments initialization or clinical trial enrollments. Rigorous evaluation of LLMs is crucial to understanding their capabilities and limitations.

Supporting information

S1 Checklist. TRIPOD checklist: Prediction model development.
(PDF)

S1 Section. Characteristics of patients with cognitive tests with and without MRI in the system, vs the subset of those with an MRI in the system.
(DOCX)

S2 Section. Examples of the notes and corresponding output that is produced by ChatGPT.(Dates for each patient note and ChatGPT responses are shifted by a random year and month, to preserve anonymous nature of notes. All note shifts for one patients are consistent).
(DOCX)

S3 Section. Parsing the JSON results.
(DOCX)

S4 Section. JSON responses.
(DOCX)

S5 Section. Qualitative analysis of error instances of ChatGPT.
(DOCX)

S1 Table. Diagnosis criteria for cognitively normal, mild cognitive impairment and mild Alzheimer's disease dementia in ADNI cohorts.
(DOCX)

S2 Table. Description of the errors encountered by ChatGPT API.
(DOCX)

S3 Table. The prompt (and the full request JSON for the task) for ChatGPT. CLINICAL_NOTE would include the date of the note (from EPIC) + ":" + the text-only content of the notes.
(DOCX)

S4 Table. The prompt (and the full request JSON for the task) for LlaMA-2. CLINICAL_NOTE would include the date of the note (from EPIC) + ":" + the text-only content of the notes.
(DOCX)

Acknowledgments

The following 22 authors are our clinical reviewers who also contributed to reviewing and authorship of the manuscript: N.G, I.S.J, A.B.C, N.H., N.F.A, L.J.B., A.J.C., Z.K., E.C.S., J. P., J. S., K.L., B.S., S.M., E.J.K., J.L., T.M.H, A.A., E.C., J.D., W.S., E.C.; Authors N.J., V.J.M. H.G., and Y.A., provided significant contributions to dataset construction, Redcap evaluation design and analysis and writing. Author Simon Jones performed statistical analysis. Authors J.A.D., A.A.B., and A.M. provided significant domain expertise in conceptualization and assistance in writing. Author N.R. led the study, assembled the team, and supervised the full execution of the study and is the corresponding author. Author H.Z. completed all Llama-2 analysis and helped in writing.

Author Contributions

Conceptualization: Narges Razavian.

Data curation: Vincent J. Major, Himanshu Grover, Yindalon Aphinyanaphongs, Narges Razavian.

Formal analysis: Hao Zhang, Simon Jones, Nicholas Genes, Ian S. Jaffe, Anthony B. Cardillo, Noah Heilenbach, Nadia Fazal Ali, Luke J. Bonanni, Andrew J. Clayburn, Zain Khera, Erica C. Sadler, Jaideep Prasad, Jamie Schlacter, Kevin Liu, Benjamin Silva, Sophie Montgomery, Eric J. Kim, Jacob Lester, Theodore M. Hill, Alba Avoricani, Ethan Chervonski, James Davydov, William Small, Eesha Chakravartty, Narges Razavian.

Funding acquisition: Narges Razavian.

Investigation: Narges Razavian.

Methodology: Narges Razavian.

Project administration: Neil Jethani, Narges Razavian.

Resources: Narges Razavian.

Software: Hao Zhang, Narges Razavian.

Supervision: Neil Jethani, John A. Dodson, Abraham A. Brody, Arjun Masurkar, Narges Razavian.

Validation: Narges Razavian.

Writing – original draft: Neil Jethani, Narges Razavian.

Writing – review & editing: Simon Jones, Nicholas Genes, Vincent J. Major, Anthony B. Cardillo, Noah Heilenbach, Nadia Fazal Ali, Luke J. Bonanni, Andrew J. Clayburn, Zain Khera, Erica C. Sadler, Jaideep Prasad, Jamie Schlacter, Kevin Liu, Benjamin Silva, Sophie Montgomery, Eric J. Kim, Jacob Lester, John A. Dodson, Abraham A. Brody, Yindalon Aphinyanaphongs, Arjun Masurkar, Narges Razavian.

References

1. OpenAI. ChatGPT. 2023 [cited 3 Jul 2023]. Available: http://openai.com/chatgpt (accessed June 2023)
2. OpenAI. GPT-4 Technical Report. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.08774
3. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2305.09617
4. Touvron H, Louis Martin, Stone K, Albert P, Almahairi A, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
5. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.12712
6. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.13375
7. Lee P, Bubeck S, Petro J, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023; 388: 1233–1239.
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023; 2: e0000198. https://doi.org/10.1371/journal.pdig.0000198 PMID: 36812645
9. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurology Open. 2023; 5. https://doi.org/10.1136/bmjno-2023-000451 PMID: 37337531
10. Matias Y. Our latest health AI research updates. In: Google [Internet]. 14 Mar 2023 [cited 3 Jul 2023]. Available: https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/
11. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023; 329: 842–844. https://doi.org/10.1001/jama.2023.1044 PMID: 36735264
12. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology. 2023; 307: e230424. https://doi.org/10.1148/radiol.230424 PMID: 37014239
13. Lee T-C, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT Answers Common Patient Questions About Colonoscopy. Gastroenterology. 2023. https://doi.org/10.1053/j.gastro.2023.04.033 PMID: 37150470
14. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023; 183: 589–596. https://doi.org/10.1001/jamainternmed.2023.1838 PMID: 37115527
15. Dash D, Thapa R, Banda JM, Swaminathan A, Cheatham M, Kashyap M, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2304.13714

16. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023; 47: 33. https://doi.org/10.1007/s10916-023-01925-4 PMID: 36869927

17. Koo M. The Importance of Proper Use of ChatGPT in Medical Writing. Radiology. 2023; 307: e230312. https://doi.org/10.1148/radiol.230312 PMID: 36880946

18. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. In: Nature Publishing Group UK [Internet]. 18 Jan 2023 [cited 4 Jul 2023]. https://doi.org/10.1038/d41586-023-00107-z PMID: 36653617

19. Thorp HH. ChatGPT is fun, but not an author. Science. 2023; 379: 313–313. https://doi.org/10.1126/science.adg7879 PMID: 36701446

20. Nature. Authorship. In: Nature Authorship [Internet]. Springer Nature; 2023 [cited 4 Jul 2023]. Available: https://www.nature.com/nature/editorial-policies/authorship

21. JAMA. Instructions for Authors. In: JAMA Authorship Guidelines [Internet]. 4 Jul 2023 [cited 4 Jul 2023]. Available: https://jamanetwork.com/journals/jama/pages/instructions-for-authors

22. Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. Account Res. 2023; 1–9. https://doi.org/10.1080/08989621.2023.2168535 PMID: 36697395

23. Park D. Open LLM Leaderboard. In: Open LLM Leaderboard [Internet]. 4 Jul 2023 [cited 4 Jul 2023]. Available: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

24. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010; 17: 229–236. https://doi.org/10.1136/jamia.2009.002733 PMID: 20442139

25. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc. 2018; 25: 331–336. https://doi.org/10.1093/jamia/ocx132 PMID: 29186491

26. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc. 2018; 25: 530–537. https://doi.org/10.1093/jamia/ocx160 PMID: 29361077

27. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010; 17: 507–513. https://doi.org/10.1136/jamia.2009.001560 PMID: 20819853

28. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. J Biomed Inform. 2001; 34: 301–310. https://doi.org/10.1006/jbin.2001.1029 PMID: 12123149

29. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. pp. 2097–2106.

30. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901 07031. 2019. Available: https://www.aaai.org/Papers/AAAI/2019/AAAI-IrvinJ.6537.pdf

31. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. arXiv [cs.CL]. 2020. Available: http://arxiv.org/abs/2004.09167

32. McDermott MBA, Hsu TMH, Weng W-H, Ghassemi M, Szolovits P. CheXpert++: Approximating the CheXpert labeler for Speed,Differentiability, and Probabilistic Output. arXiv [cs.LG]. 2020. Available: http://arxiv.org/abs/2006.15229

33. Le Glaz A, Haralambous Y, Kim-Dufor D-H, Lenca P, Billot R, Ryan TC, et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. J Med Internet Res. 2021; 23: e15708. https://doi.org/10.2196/15708 PMID: 33944788

34. Weng W-H, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak. 2017; 17: 1–13.

35. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023; 1–6. https://doi.org/10.1038/s41586-023-06160-y PMID: 37286606

36. Leiter RE, Santus E, Jin Z, Lee KC, Yusufov M, Chien I, et al. Deep Natural Language Processing to Identify Symptom Documentation in Clinical Notes for Patients With Heart Failure Undergoing Cardiac Resynchronization Therapy. J Pain Symptom Manage. 2020; 60: 948–958.e3. https://doi.org/10.1016/j.jpainsymman.2020.06.010 PMID: 32585181

37. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc. 2015; 23: e20–e27. https://doi.org/10.1093/jamia/ocv130 PMID: 26338219

38. Taggart M, Chapman WW, Steinberg BA, Ruckel S, Pregenzer-Wenzler A, Du Y, et al. Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. JAMA Netw Open. 2018; 1: e183451–e183451. https://doi.org/10.1001/jamanetworkopen.2018.3451 PMID: 30646240

39. Wu Y, Denny JC, Trent Rosenbloom S, Miller RA, Giuse DA, Xu H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. AMIA Annu Symp Proc. 2012; 2012: 997. PMID: 23304375

40. Fan Y, Wen A, Shen F, Sohn S, Liu H, Wang L. Evaluating the Impact of Dictionary Updates on Automatic Annotations Based on Clinical NLP Systems. AMIA Summits Transl Sci Proc. 2019; 2019: 714. PMID: 31259028

41. Larochelle H, Erhan D, Bengio Y. Zero-data learning of new tasks. Proceedings of the 23rd national conference on Artificial intelligence—Volume 2. AAAI Press; 2008. pp. 646–651.

42. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Finetuned language models are zero-shot learners. arXiv [cs.CL]. 2021. Available: https://research.google/pubs/pub51119/

43. Rezaei M, Shahidi M. Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review. Intelligence-Based Medicine. 2020; 3–4: 100005. https://doi.org/10.1016/j.ibmed.2020.100005 PMID: 33043311

44. Borji A. A Categorical Archive of ChatGPT Failures. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2302.03494

45. Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 1906–1919.

46. Tsoy E, Kiekhofer RE, Guterman EL, Tee BL, Windon CC, Dorsman KA, et al. Assessment of Racial/Ethnic Disparities in Timeliness and Comprehensiveness of Dementia Diagnosis in California. JAMA Neurol. 2021; 78: 657–665. https://doi.org/10.1001/jamaneurol.2021.0399 PMID: 33779684

47. Lin P-J, Daly A, Olchanski N, Cohen JT, Neumann PJ, Faul JD, et al. Dementia diagnosis disparities by race and ethnicity. Alzheimers Dement. 2020; 16. https://doi.org/10.1002/alz.043183

48. Saadi A, Himmelstein DU, Woolhandler S, Mejia NI. Racial disparities in neurologic health care access and utilization in the United States. Neurology. 2017; 88: 2268–2275. https://doi.org/10.1212/WNL.0000000000004025 PMID: 28515272

49. Drabo EF, Barthold D, Joyce G, Ferido P, Chang Chui H, Zissimopoulos J. Longitudinal analysis of dementia diagnosis and specialty care among racially diverse Medicare beneficiaries. Alzheimers Dement. 2019; 15: 1402–1411. https://doi.org/10.1016/j.jalz.2019.07.005 PMID: 31494079

50. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet. 2020; 396: 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6 PMID: 32738937

51. Harper LC. 2022 Alzheimer's Association Facts and Figures. https. Available: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/915A476B938D0AF39A218D34852AF645/9781009325189mem_205-207.pdf/resources.pdf

52. US Dept of Health and Human Services. National Plan to Address Alzheimer's Disease: 2020 Update. 2021 [cited 1 Nov 2021]. Available: https://aspe.hhs.gov/reports/national-plan-address-alzheimers-disease-2020-update-0

53. SPRINT MIND Investigators for the SPRINT Research Group, Williamson JD, Pajewski NM, Auchus AP, Bryan RN, Chelune G, et al. Effect of Intensive vs Standard Blood Pressure Control on Probable Dementia: A Randomized Clinical Trial. JAMA. 2019; 321: 553–561. https://doi.org/10.1001/jama.2018.21442 PMID: 30688979

54. Pragmatic Evaluation of Events And Benefits of Lipid-lowering in Older Adults—Full Text View—ClinicalTrials.Gov. [cited 27 Oct 2021]. Available: https://clinicaltrials.gov/ct2/show/NCT04262206

55. NIA. NIA-funded active Alzheimer's and related dementias clinical trials and studies. In: NIA [Internet]. 2021 [cited 20 Apr 2021]. Available: https://www.nia.nih.gov/research/ongoing-AD-trials

56. Science. In: AAAS [Internet]. [cited 10 Jul 2023]. Available: https://www.science.org/content/article/another-alzheimers-drug-flops-pivotal-clinical-trial

57. Drug Approval Package: Aduhelm (aducanumab-avwa). [cited 31 Oct 2021]. Available: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/761178Orig1s000TOC.cfm

58. Manly JJ, Glymour MM. What the Aducanumab Approval Reveals About Alzheimer Disease Research. JAMA Neurol. 2021. https://doi.org/10.1001/jamaneurol.2021.3404 PMID: 34605885

59. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State Examination. J Psychiatr Res. 1975. https://doi.org/10.1037/t07757-000

60. Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology. 1993. pp. 2412–2412. https://doi.org/10.1212/wnl.43.11.2412-a PMID: 8232972

61. Collins G.S., Reitsma J.B., Altman D.G. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 13, 1 (2015). https://doi.org/10.1186/s12916-014-0241-z PMID: 25563062

62. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005; 53: 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x PMID: 15817019

63. ADNI. 2021 [cited 1 Nov 2021]. Available: http://adni.loni.usc.edu/data-samples/adni-participant-demographic/

64. Azure OpenAI Service content filtering—Azure OpenAI. [cited 10 Jul 2023]. Available: https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/content-filter

65. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971; 76: 378–382.

66. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012; 8: 23. https://doi.org/10.20982/tqmp.08.1.p023 PMID: 22833776

67. Maxwell AE. Coefficients of Agreement Between Observers and Their Interpretation. Br J Psychiatry. 1977; 130: 79–83. https://doi.org/10.1192/bjp.130.1.79 PMID: 831913

68. Beekly DL, Ramos EM, van Belle G, Deitrich W, Clark AD, Jacka ME, et al. The National Alzheimer's Coordinating Center (NACC) Database: an Alzheimer disease database. Alzheimer Dis Assoc Disord. 2004; 18: 270–277. PMID: 15592144