

Leveraging large language model (LLM)-powered agents when addressing problems in mental health and chronic conditions: a scoping review

Qingfan An, Po-Yin Yen

Submitted to: Journal of Medical Internet Research
on: April 03, 2025

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 its 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 expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 26

..... 27

Multimedia Appendixes 28

 Multimedia Appendix 1..... 29

CONSORT (or other) checklists..... 30

 CONSORT (or other) checklist 0..... 30

Leveraging large language model (LLM)-powered agents when addressing problems in mental health and chronic conditions: a scoping review

Qingfan An¹; Po-Yin Yen² PhD, FAAN

¹ Umeå University Umeå SE

² Washington University School of Medicine in St Louis St Louis US

Corresponding Author:

Qingfan An

Umeå University

Biologihuset Umeå universitet, 901 87 Umeå

Umeå

SE

Abstract

Background: Recent advancements in large language models (LLMs) present promising opportunities for improving the management of chronic health conditions and mental health challenges. However, their integration into healthcare remains complex, especially with rapid developments like ChatGPT since 2022. These developments highlight the need for collaboration between healthcare professionals (HCPs) and LLM-powered agents to ensure safe and effective adoption.

Objective: This scoping review aims to assess and synthesize how LLM agents are integrated into healthcare to support chronic disease and mental health management while guiding informed adoption.

Methods: We conducted a scoping review following Arksey and O'Malley's framework. We searched PubMed/MEDLINE, Embase, ACM Digital Library, and Web of Science for English-language articles published between 1st Jan 2020 to 17th April 2024 that examined the use of LLM-powered agents as assistive tools in managing chronic conditions and mental health. After screening for inclusion, we performed data extraction and summarized the findings.

Results: We total included 14 studies. Among those six used pretrained models without any modification, while eight studies fine-tuned or augmented the models to improve performance. LLM-powered agents made contributions across various medical decision-making scenarios by assisting with health data interpretation, automating clinical assessment and monitoring, and providing clinical decision references. Three key aspects of LLM evaluation emerged: model performance, comparative evaluation, and human evaluation.

Conclusions: Responsible integration of LLM-powered agents requires caution due to early-stage challenges like data limitations and AI hallucinations. Sustainable adoption will benefit from practical insights into user interactions and real-world implementation. Future studies should establish performance benchmarks, focus on user experience, improve LLM functionality, and set protocols for long-term evaluation, addressing stability and reliability to ensure lasting benefits in patient care.

(JMIR Preprints 03/04/2025:75431)

DOI: <https://doi.org/10.2196/preprints.75431>

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 visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/75431>, I will be able to make my manuscript PDF available to anyone. No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

Leveraging large language model (LLM)-powered agents when addressing problems in mental health and chronic conditions: a scoping review

Abstract

Background: Recent advancements in large language models (LLMs) present promising opportunities for improving the management of chronic health conditions and mental health challenges. However, their integration into healthcare remains complex, especially with rapid developments like ChatGPT since 2022. These developments highlight the need for collaboration between healthcare professionals (HCPs) and LLM-powered agents to ensure safe and effective adoption.

Objective: This scoping review aims to assess and synthesize how LLM agents are integrated into healthcare to support chronic disease and mental health management while guiding informed adoption.

Methods: We conducted a scoping review following Arksey and O'Malley's framework. We searched PubMed/MEDLINE, Embase, ACM Digital Library, and Web of Science for English-language articles published between 1st Jan 2020 to 17th April 2024 that examined the use of LLM-powered agents as assistive tools in managing chronic conditions and mental health. After screening for inclusion, we performed data extraction and summarized the findings.

Results: We total included 14 studies. Among those six used pretrained models without any modification, while eight studies fine-tuned or augmented the models to improve performance. LLM-powered agents made contributions across various medical decision-making scenarios by assisting with health data interpretation, automating clinical assessment and monitoring, and providing clinical decision references. Three key aspects of LLM evaluation emerged: model performance, comparative evaluation, and human evaluation.

Conclusions: Responsible integration of LLM-powered agents requires caution due to early-stage challenges like data limitations and AI hallucinations. Sustainable adoption will benefit from practical insights into user interactions and real-world implementation. Future studies should establish performance benchmarks, focus on user experience, improve LLM functionality, and set protocols for long-term evaluation, addressing stability and reliability to ensure lasting benefits in patient care.

Keywords: large language models; LLM; chronic disease; mental health; ChatGPT; scoping; review; NLP; chronic conditions; Natural Language Processing

Introduction

Recent advancements in large language models (LLMs) have provided substantial benefits for tasks requiring sophisticated text processing and generation capabilities. These advancements offer promising opportunities for healthcare across various domains, including research, education, and patient care [1], particularly in addressing complex areas such as chronic health conditions and mental health challenges [2-4].

Chronic conditions, which include four main types, cardiovascular diseases, cancer, diabetes, and chronic respiratory diseases, demand long-term management that can span years or even decades. These conditions account for nearly 70% of deaths worldwide and an increasing burden of morbidity that reduces productivity, increases medical spending, and shortens lives [5-7]. Unlike acute conditions, chronic illnesses necessitate lifelong treatment, placing substantial demands on healthcare resources. The economic impact is significant, with chronic diseases consuming up to 7% of a nation's GDP due to their negative effects on economic activities and the associated rise in

public health and social welfare expenditures [8-9]. The main cost drivers in healthcare systems include increased service utilization, medication expenses, and the need to adapt health services [8]. Artificial Intelligence (AI) has shown promise in reducing such costs associated with chronic care [10]. Within AI, Natural Language Processing (NLP) stands out as particularly beneficial. NLP has proven valuable in various aspects of chronic disease management such as diagnosis [11], interpretation of clinical narratives [12], and patient self-management [13]. Chronic health conditions often intersect with mental health challenges, with individuals suffering from chronic diseases being two to three times more likely to experience concurrent mental health issues compared to the general population [14]. When mental health issues persist, they can also be classified as chronic conditions, requiring attentive, responsive care with a focus on increasing awareness and encouraging proactive help-seeking. NLP methods offer valuable insights in addressing mental health including extracting symptoms, classifying severity of illness, comparing therapy effectiveness, providing psychopathological clues, and challenging the current nosography [15].

LLM-powered agents, as a specialized development within NLP, have the potential to become crucial partners in healthcare delivery, enhancing efficiency and reducing long-term costs [16]. However, responsibly integrating these technologies poses significant challenges, particularly in light of developments like ChatGPT, which has spurred rapid progress since 2022. The increasing capabilities of LLM-powered agents introduce new complexities and concerns, especially when tasked with interpreting vast amounts of clinical data to provide patient-specific insights [1-3]. These models are not yet ready to serve as standalone decision-making tools due to issues such as AI hallucinations, ethical concerns, and the need for greater accountability [17-21]. This underscores the necessity for collaboration between healthcare professionals (HCPs) and these agents, emphasizing the need to involve professional expertise in navigating these advancements. Studies have demonstrated the potential of such partnerships in adapting clinical guidelines [22], medical decision making [23], and patient education [24]. However, these collaboration modes are diverse and dynamic, constrained by existing healthcare paradigms. Given the rigid nature of these paradigms, further exploration of collaboration modes and recommendations for improvement could facilitate more effective partnerships in the future, particularly in managing chronic diseases and mental health care. Our focus is on ensuring sustainable and responsible deployment to maximize the benefits of LLM-powered agents in clinical practice. This review aims to explore existing collaborations between LLM-powered agents and stakeholders in the healthcare environment for managing chronic conditions and mental health care. Our research question is: "How LLM-powered agents are employed as assistive tools in to enhance the management of mental health and chronic condition challenges?" We examined the research studies for the training and model refinement strategies utilized in developing these models and their practical applications in clinical practice, and evaluated their performance, along with the methods used to assess their performance in studies where they have been employed as assistive tools.

Methods

This scoping review was conducted following the Arksey and O'Malley framework [25] and recommendations for clarification of the Arksey and O'Malley framework proposed by Levac et al [26]. It is reported according to the PRISMA Extension for Scoping Reviews (PRISMA-ScR) checklist (Multimedia Appendix 1). The protocol was registered at Open Science Framework (Multimedia Appendix 2). The search terms remained unchanged as they were in the registered protocol.

Search strategy

A literature search was performed using PubMed/MEDLINE, Embase, ACM digital library, and Web of Science databases. We consulted a medical librarian to develop the literature search strategy. The

search was conducted using two keyword groups: "mental health" "chronic disease" and "LLM". Multimedia Appendix 3 contains the full search strategy for four databases.

Study screening

Studies were screened according to a three-step process that encompassed i) managing search results and removing duplicates using Mendeley and Rayyan; ii) title and abstract screening; and iii) full-text screening. This scoping review focused on English-language studies published between 1st Jan 2020 to 17th April 2024 that evaluated the use of LLM-powered agents as assistive tools in managing chronic diseases and mental health conditions. Two reviewers (QA, PY) independently conducted the title and abstract screening to assess the initial pool of identified citations for eligibility. This was followed by an independent full-text screening by the same reviewers (QA, PY) to confirm study inclusion. Any conflicts arising during either the title/abstract or full-text screening stages were resolved through discussion between the reviewers. Additionally, two reviewers (QA, PY) independently screened titles and abstracts from relevant grey literature, which was identified through a manual search. The screening process is visualized in the PRISMA flow diagram in Fig. 1. The specific inclusion and exclusion criteria applied during the screening process are detailed in Table 1.

Table 1. Inclusion and exclusion criteria.

Inclusion criteria	
1	Articles evaluated the LLMs in treating chronic conditions or mental health problems in English.
2	Original peer-reviewed articles published as a journal article
4	The use of LLM in addressing chronic conditions or mental health problems is the primary subject of this study. (Mental health problems examined do not need to be chronic)
5	HCPs remain in control of the care process, with LLM-powered agents serving in an assistive role explicitly mentioned in the report. Articles attempting to replace HCPs with ChatGPT are excluded.
Exclusion criteria	
1	Review articles, protocols, and letters to editors are excluded.
2	Interventions aimed at detecting, diagnosing, or preventing chronic diseases or mental health problems are excluded.
3	Articles without evaluation reports are excluded.
4	Articles that focus on mental health issues or chronic diseases but also include other conditions with comparable symptoms are excluded.
5	Articles not specific to chronic conditions are excluded (e.g., only cardiovascular without mentioning if it is chronic).
6	Qualitative studies that investigate broad perspectives on LLMs are excluded (only studies on specific LLMs are included).

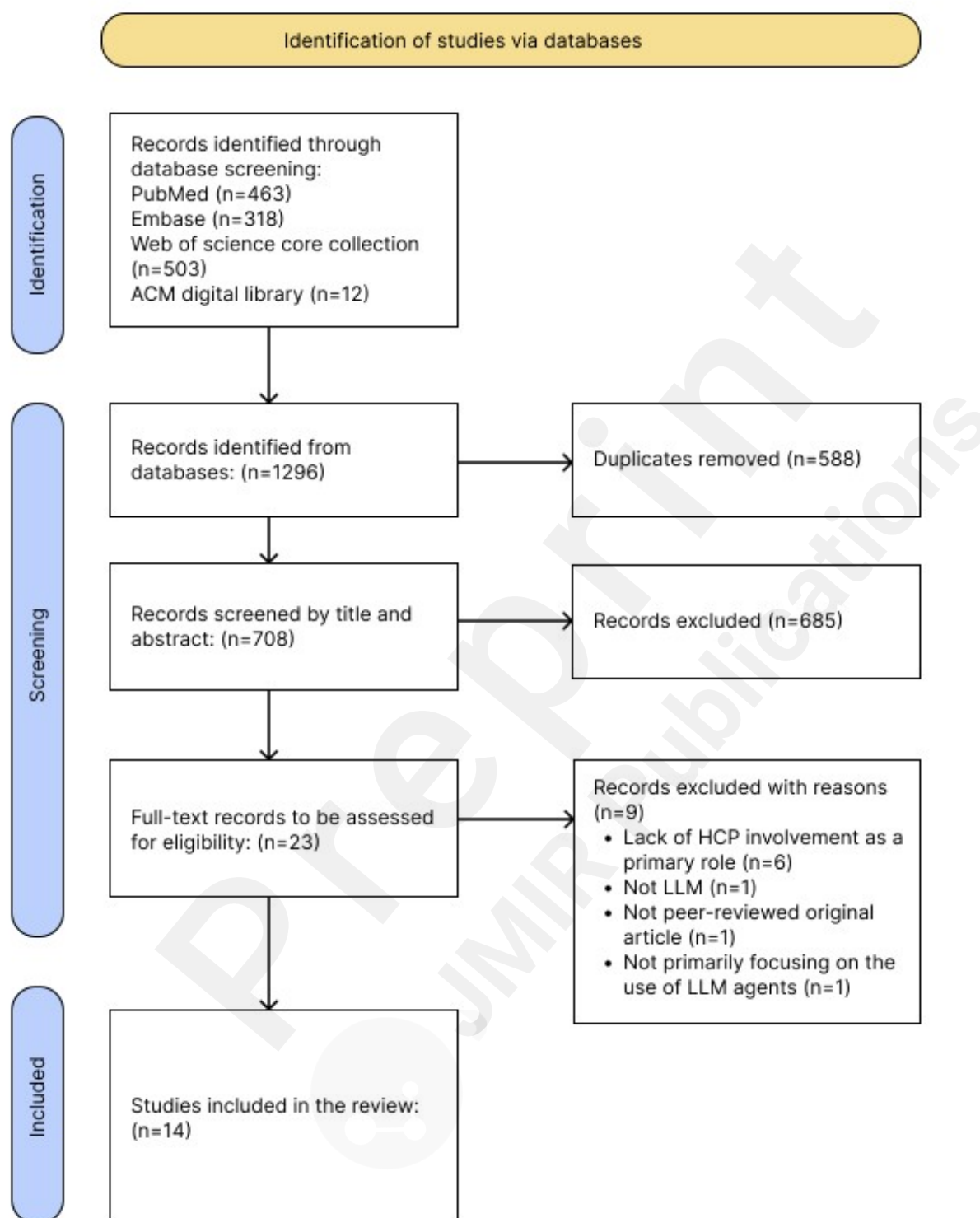


Figure 1. PRISMA Flow Diagram for the scoping review process.

Data extraction

Data extraction was carried out independently by the primary author (Q.A.), who documented the information in a Microsoft Excel spreadsheet. The extracted data encompassed various aspects: study characteristics, strategies for LLM training and refinement, the application of LLMs in clinical settings, their performance across different use cases, and the evaluation methods used to assess their effectiveness. The extraction table was subsequently validated by the co-author (P.Y.) to ensure accuracy. In addition to the tabulated data, a narrative summary was provided. This summary connects the findings to the overall objectives and the research question of the review and offers insights into how these results can inform the future integration of LLM-powered agents in the management of mental health and chronic conditions.

Results

14 studies were finally included from 1296 retrieved citations from four databases (Figure1).

Study characteristics

All 14 studies included were published between 2023 and 2024. The characteristics of these studies (year, country, focused conditions, primary user of the agent, study aim and study design) are detailed in Table 2. Four were conducted in Europe, six in the United States, and four in Asia. Of these, two studies focused on mental health problems [27 28] , while the remaining 12 articles addressed various chronic conditions [29-40]. Among the mental health studies, one used LLMs to assist in assessing depression severity [27], and the other supported psychotherapists in daily consultations [28]. The chronic condition studies predominantly concentrated on cancer patients, covering a range of types such as lung cancer [29-31], head and neck cancer [32], prostate cancer [30 33 34], breast cancer [29 30 35], gynecological cancer [29], colorectal cancer [29] and thyroid cancer [36 37]. Other chronic conditions examined included diabetes [38], stroke [39], and hematologic malignancies [40]. Most studies employed a methodological, observational, or experimental study design, primarily dealing with quantitative data [27 29-32 34-40], with one being a qualitative analysis [28]. The primary objectives of the included studies were to evaluate the feasibility of using specific LLM-powered agents in distinct scenarios [27 28 30 32 34-36 38-40], compare different interaction patterns of LLM agent use [33], or assess the performance of various LLM-powered agents across specific contexts [27 29 31 37].

Table 2. Studies evaluating LLMs for addressing chronic conditions and mental health issues

Year, Country, Author	Chronic condition or mental health issues	Primary user of the agent	Study aim	Study design
2023, Korea, Son [30]	Consecutive cancer	Nuclear medicine specialists	To investigate the feasibility of developing a deep learning model for detecting bone metastasis in bone scans using ChatGPT.	An observational diagnostic accuracy study
2024, China, Gu [39]	Stroke	Stroke specialists	To develop an automatic, quantitative stroke	An algorithm development and validation study

			severity assessment framework through automating the entire National Institutes of Health Stroke Scale (NIHSS) scoring process on Chinese clinical electronic health records (EHRs).	(methodological study)
2023, Germany, Fink [31]	Lung cancer	Lung cancer radiologists	To compare the performance of the ChatGPT and GPT-4 in data mining and labeling oncologic phenotypes from free-text CT reports on lung cancer by using user-defined prompts.	A retrospective comparative performance evaluation study (methodological study)
2024, Belgium. Lechien [32]	Head and neck cancer	Head and neck cancer oncologists	To evaluate the ChatGPT-4 performance in oncological board decisions.	A prospective observational study
2023, USA, Wu [33]	Prostate cancer	Prostate radiation oncologists	To explore using GPT-4 for radiation toxicity monitoring in prostate cancer treatments using two different methods: a summarization method and a chatbot interface.	A prospective observational study
2024, USA, Pakhomov [40]	Hematologic malignancies	Patients with hematologic malignancies	To introduce and evaluate the Automated Neural Nursing Assistant (ANNA) system, which is designed for intensive monitoring of cognitive symptoms associated with neurotoxicity following immunotherapy of hematologic	A pilot study and a prospective observational study

			malignancies.	
2023, Romania, Albajy [38]	Diabetes	Diabetes specialists	To develop predictive models for High-Density Lipoprotein levels in diabetic patients using various regression techniques utilizing multiple regression techniques with the assistance of ChatGPT's Code Interpreter plugin.	A prospective observational study design combined with machine learning analysis
2023, Singapore, Tan [29]	Cancer	Cancer specialists, particularly oncologists or radiologists	To evaluate LLMs' ability to accurately infer cancer disease response from free-text radiology reports.	A retrospective observational study
2023, USA, Gebrael [34]	Metastatic prostate cancer	Emergency room HCPs	To investigate the effectiveness of ChatGPT, an advanced artificial intelligence system, in assisting health providers with decision-making for patients presenting with metastatic prostate cancer in the emergency room.	A retrospective observational study
2023, USA, Lian [36]	Thyroid cancer	Thyroid cancer specialists	To create a preliminary NLP pipeline for extracting health-related quality of life (HRQOL) trajectories from patient language, utilizing ChatGPT for data augmentation.	A methodological study
2023, Korea, Choi [35]	Breast cancer	Breast cancer specialists	To evaluate the time and cost of developing prompts using LLM, tailored to extract clinical factors in breast	A retrospective observational study

			cancer patients and their accuracy.	
2023, USA, Pathak [37]	Thyroid cancer	Thyroid specialists	To improve the extraction of thyroid nodule characteristics from ultrasound reports using LLMs.	A methodological study
2023, USA, Lau [27]	Depression	Psychologists working with depression	To tackle the issue of data scarcity in depression assessment by employing LLMs for data augmentation.	An experimental study
2024, Austria, Raile [28]	Psychological issues	Patients with psychological issues	To evaluate the role and effectiveness of ChatGPT as a supplementary resource for mental health support.	A qualitative analysis

LLM training and model refinement

Table 3 provides detailed descriptions of the adaptation of LLMs in the included studies. All included studies adapted LLMs with ChatGPT being the most frequently utilized. Most studies did not provide explicit reasons for their choices of the initial LLM, although one study justified their choice influenced by similar usage in other research [28]. Six studies directly used the pretrained models without modification, whereas eight studies fine-tuned the pretrained LLMs or conducted data augmentation for training data to improve the model performance. Among included architectures, the GPT architecture was the most commonly chosen for those model refinement.

The datasets used for fine-tuning were diverse and tailored to the specific needs of each study. These included electronic health records (EHRs) [38 39], free-text radiology reports [29 31 35 37], and patient interview data [27 36]. Some studies also incorporated data augmentation techniques to enrich their training datasets. For instance, Tan et al. used sentence permutation to create synthetic radiology reports, thereby expanding the training data and increasing the sample size, which led to improved accuracy [29]. They also demonstrated that the ADAPET method, a prompt-based fine-tuning approach using the GatorTron model, effectively reduced the required number of training reports [29]. Similarly, Lian utilized GPT-2 for synthetic data generation alongside a BERT-based model, yielding promising results in predicting health-related quality of life (HRQOL) [36]. Lau et al., introduced an ultra-lightweight alternative to traditional fine-tuning known as prefix-tuning [27]. Prefix-tuning decouples the number of tuning parameters from the model architecture, allowing for flexible adjustments to the model's learning capacity [27]. The prefix-tuned model outperformed all previously published methods [27].

The outputs generated by these LLM-powered agents to assist in decision-making were diverse, with the goal of streamlining workflows and enhancing patient care. The outputs included advanced deep learning models for medical image analysis and other clinical data analysis [30] [38], automated assessments of patients' cognitive function and condition severity [27 36 39 40], extraction of clinical information [31] [32] [35] [37], data augmentation and prompting [29], triage suggestions [34], narratives for psychological consultation [28], and summaries based on patient-reported

outcomes [33].

Despite these promising advancements in different contexts, several limitations were identified in the adaptation of LLM agents across the studies. One notable limitation was the small size of training datasets, which potentially impacted the performance and accuracy of the models [39]. Additionally, authors claimed that the reliance on closed-source, commercial LLMs posed challenges regarding transparency and adaptability [33]. AI hallucination was mentioned as well when models like ChatGPT produce outputs that seems plausible but were factually incorrect or irrelevant to the input context [34] [40]. Stability concerns were also observed with ChatGPT-4 over time [32], and some studies reported issues with response latency, affecting the practical deployment of these models in real-time clinical environments [40].

Table 3. Adaptation details of LLMs.

Year, Country, Author	Platform	Model	Was the LLM refined, If so, what data/method used for refinement	Outputs
2023, Korea, Son [30]	ChatGPT 3.5	GPT-3.5	No	Deep learning model for medical image analysis
2024, China, Gu [39]	LLM-based CNER (Clinical Named Entity Recognition) system	CliRoberta	Trained using domain-adaptive transfer learning on Chinese clinical EHRs data.	Severity scores for stroke
2023, Germany, Fink [31]	ChatGPT	GPT-3.5 / GPT-4	Task-specific training or instruction for the LLMs	Extraction of oncologic information
2024, Belgium, Lechien [32]	Chat GPT	GPT-4	No	Extraction of oncologic information
2023, USA, Wu [33]	Chat GPT	GPT-4	No	Patient summary
2024, USA, Pakhomov [40]	Automated Neural Nursing Assistant (ANNA)	LLaMA 2, Vicuna, Blenderbot	No	Automated assessments of their cognitive function
2023, Romania, Albajy	ChatGPT	GPT-3.5	No	Predictive models for data analysis

[38]				
2023, Singapore, Tan [29]	AI system for cancer disease response assessment	GatorTron	Cancer radiology reports for data augmentation	Data augmentation and prompting
2023, USA, Gebrael [34]	ChatGPT	GPT-4	No	Suggestions for triage
2023, USA, Lian [36]	The NLP pipeline built using BERT and supplemented with GPT-2	BERT	Interviews of 100 patients undergoing surgical intervention for low-risk thyroid cancer, paired with HRQOL assessments completed during the same visits.	Quality of life
2023, Korea, Choi [35]	GPT for Sheets and Docs extension plugin	GPT-3.5 / GPT-4	GPT for Sheets and Docs extension plugin to train the LLM.	Extraction of clinical information
2023, USA, Pathak [37]	Hugging Face's Transformers library	GatorTron	Fine-tuning the model on annotated ultrasound reports to optimize its performance for this specific task.	Extraction of clinical information
2023, USA, Lau [27]	Deep learning framework (not specified)	BERT / RoBERTa	The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ).	Severity scores for depression
2024, Austria, Raile [28]	ChatGPT	GPT-4	No	Narratives for psychological consultation

LLMs in supporting clinical practice

LLMs in supporting clinical practice LLM-powered agents made contributions across various medical decision-making scenarios by assisting with health data interpretation, automating clinical assessment and monitoring, and supplying clinical decision support references.

By assisting in the interpretation of complex scientific data and improving predictive accuracy, LLM agents reduce the programming and data analysis burden on HCPs. They supported HCPs by developing advanced models, such as those used in medical image analysis and factor analysis [30 38]. Interestingly, in one study focused on constructing medical image analysis models using LLM agents, a strategy emerged with the utilization of heatmaps generated by GPTs as well [30]. The class activation map algorithm produced visual representations that highlighted key areas of interest during bone scan evaluations for metastasis, thus improving user interaction and interpretation [30]. However, challenges remain when LLM assisting with health data interpretation, including occasional inaccuracies in ChatGPT's responses and difficulties in ensuring error-free code [30].

During clinical visits, LLM agents have strengthened the process of clinical assessment and symptom monitoring. They facilitated triage decisions and improved accuracy in emergency rooms by analyzing patient outcomes and resource allocation [34], and automated symptom assessment and monitoring [27 33 36 39 40]. Protocol-based assessments assisted by LLMs showed promise in evaluating quality of life (QoL) depression severity and stroke severity, summarizing patient symptoms, and detecting neurotoxic effects of medications early. Notably, a novel approach for assessing depression severity has emerged, utilizing both pretrained large language models and parameter-efficient tuning techniques to overcome challenges related to data scarcity [27].

In addition, LLMs have provided valuable clinical decision support references by streamlining data processing and refining treatment recommendations [28 29 31 32 35 37]. They aided in suggesting additional examinations, management, and therapeutic approaches [32], providing automated second opinions on disease progression from radiology reports [29], data mining and labeling oncologic phenotypes from free-text lung cancer CT reports [31], and extracting thyroid nodule characteristics from ultrasound reports to balance the daily practice report quality [37]. They were proved efficiently extract clinical features, covering more critical aspects than traditional manual methods or standard language models [31 35 37]. ChatGPT has also been beneficial in psychotherapeutic settings, praised for its cost-free accessibility, widespread use, interactivity, empathy, and overall contribution to mental health treatment [28]. Additionally, Choi et al. highlighted effective prompt engineering when extracting the required information from surgical pathology reports and preoperative ultrasound reports of breast cancer patients, and stated that understanding LLM features, and avoiding jargon are crucial for desired outcomes [35].

To operate effectively in various contexts, the types of agents studied encompassed five conversational agents providing clinical advice and computational support [28 30 32 33 40], two medical text analysis agents extracting clinical details from narratives [29] [31], two severity assessment agents enabling automated severity evaluations of conditions [27 39], one code interpretation agent assisting with data analysis [38], two medical information extraction agents specifically focused on extracting clinical information from ultrasound reports [35 37], and two medical classification agents classifying HRQOL trajectories and offering triage suggestions [34 36]. While most studies primarily targeted HCPs as the primary users of LLM agents, patients were included as the main users in two studies [28 40], with no direct interaction between HCPs and LLMs. One study evaluated the Automated Neural Nursing Assistant (ANNA) system, which monitored cognitive symptoms related to neurotoxicity following immunotherapy for hematologic malignancies, allowing patients to interact with the system via phone calls [40]. Another study highlighted ChatGPT as a valuable adjunct to psychotherapy, offering an easily accessible and currently free resource for individuals with mental health issues who have not yet sought professional help [28].

Evaluation of LLM Agents

Our analysis identified three primary aspects of evaluating LLM agents as assistive tools: model performance, comparative evaluation, and human evaluation. All the studies we reviewed addressed at least one of these aspects. However, several challenges emerged during the evaluation process, including small sample sizes [28 29 31-33 36 39] and incomplete health data for evaluation [33].

Model Performance

A majority of studies focused on evaluating model's overall performance, which included metrics such as accuracy [27-29 31 32 34 35 37 39], sensitivity [31], specificity [31], Area Under the Curve (AUC) [36], hypothesis testing, stability [32], confabulation rate [31], and latency of system response [40].

Accuracy was the most frequently evaluated dimension, with generally positive reports. Specific measurements used included accuracy rate [29], F1 score [37 39], Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) [27], and accuracy questionnaires completed by HCPs [33] and Amazon Mechanical Turk workers [40]. Only one article concluded that overall accuracy remained modest and not yet ready for clinical implementation [36]. F1 scores ranged from 0.8851 to 0.990 [37 39]. Consistency of the outputs reported moderate to high consistency [32]. Despite these positive results, some stability issues were identified, particularly in providing consistent treatment recommendations for cancer patients over time [32], generating responses in psychological therapy [28].

Two studies evaluated the performance of models created by LLM-driven agents, offering indirect insights into LLM performance [30 38]. These studies showcased how LLMs helped HCPs overcome computational challenges. One study assessed a deep learning model for diagnosing bone metastasis using AUC, sensitivity, and specificity, and compared various data augmentation strategies. The model showed promising results in detecting bone metastases across diverse malignancies [30]. The other study focused on a model developed by the ChatGPT Code Interpreter to analyze complex relationships among twelve parameters in Romanian diabetic patients, using Mean Squared Error (MSE) for predictive accuracy and model comparisons, yielding positive outcomes [38].

Comparative Evaluation

Many studies directly compared the LLM-powered agents with criteria like expert opinion (ground truth) and model performance benchmarks set by other studies. Beyond performance comparisons, the key dimensions included comprehensiveness of results compared to ER physicians for triage advices [34], cost [35], and time saved for HCPs [33 35 39].

Four studies compared different models in specific scenarios. In extracting thyroid nodule characteristics, GatorTron was the best-performing model compared with BERT, RoBERTa, LongFormer, DeBERTa [37]. GPT-4 outperformed ChatGPT in data mining oncologic phenotypes from free-text CT reports on lung cancer and demonstrated better oncologic reasoning with fewer confabulations [31]. A prefix-only model outperformed BERT-PT, RoBERTa-PT, and ST-only in automatic depression assessment [27]. A discriminative Chinese clinical LLM named CliRoberta outperformed the existing general and medical Chinese LLMs in automatic assessment of stroke severity [39].

Wu conducted a comparative study of two interaction methods assisted by GPT-4 for symptom assessment in patients undergoing radiation therapy [33]. The study found that a summarization method using GPT-4, based on Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) guidelines, was more accurate and preferred over a custom GPT-4 chatbot for assessing symptoms. Both methods, however, were effective in saving approximately four minutes of clinical time per patient interaction [33].

Human Evaluation

Human evaluation concentrated on assessing interactions and validating outcomes. In one study involving the use of LLMs for monitoring radiation therapy toxicity, ten radiation oncologists rated simulated patient responses for both methods—summarization and chatbot—on a scale of 1 to 10 for accuracy, likelihood of adoption, and estimated time saved, using the Wilcoxon signed-rank test for comparison [33]. On the other hand, most outcome validation studies employed quantitative methods; only one qualitative study explored the therapeutic potential of integrating ChatGPT into psychological treatment through case studies [28]. Among the quantitative studies, five specifically recruited participants to evaluate agent performance. One study involved 10 Amazon Mechanical Turk workers who interacted with the ANNA system and subsequently completed a structured survey [40]. The findings from this study provided insights into user perceptions and interactions with AI-driven systems in simulated tasks [40]. In a different study, four radiologists evaluated the oncologic reasoning of the two LLMs, using a free-form explanation of the chosen outcome category rated for factual correctness and accuracy, based on a five-point Likert scale based on predefined criteria [31]. Additionally, two head and neck surgeons utilized the Artificial Intelligence Performance Instrument (AIPI)—a validated tool for assessing chatbot performance in medical contexts—to evaluate ChatGPT-4 [32]. The study revealed that ChatGPT-4 often suggested a greater number of additional examinations for oncological check-ups compared to practitioners on the oncological board, highlighting its proactive approach in clinical decision-making [32]. Furthermore, three stroke specialists and three trained volunteers were recruited as separate comparison groups for stroke severity assessment. An encouraging finding from this study was that it enabled novice volunteers, lacking prior clinical knowledge, to quickly learn and perform the labor-intensive and time-consuming manual assessments of stroke severity [39].

Discussion

Our review identified four critical insights that highlight both the limitations and opportunities of integrating LLM agents in addressing problems in mental health and chronic conditions. First, the current reporting on the use of LLM agents as support tools for managing chronic conditions and mental health issues lacks universally recognized metrics, particularly in the areas of model selection, refinement, evaluation, and reporting. Additionally, there is a gap in studies assessing the long-term efficacy of these agents. Second, there is a deficiency in user experience data regarding user interactions in different platforms and the practical implementation of LLM agents. Third, LLMs have shown promise in assisting HCPs not only in developing data analysis skills but also in enhancing user interaction experiences with AI-generated results. These models can provide detailed explanations when needed, improving both the interpretability and contextual understanding of complex clinical data. Fourth, the integration of LLM agents into healthcare must carefully consider challenges such as AI hallucinations, stability issues, and other factors that may compromise the reliability of their outputs. These concerns indicate that a cautious and prolonged evaluation period is necessary before these technologies can be widely adopted in clinical practice. This section will delve deeper into these four insights.

Need for Standardization

Our review highlights the need for standardized report in the use and evaluation of LLM agents. Specifically, the selection of pre-trained models requires more explicit rationale, given the variety of LLMs available and the fact that the optimal choice may vary depending on the study context. Many studies fail to justify their model selection, raising an important question: “How should one choose the most appropriate model for addressing chronic diseases and mental health challenges?” This question becomes even more complex for healthcare professionals, who must navigate both clinical

and technical considerations in their decision-making process. As of June 2023, over 20 LLMs are available for public use, with OpenAI's GPT-4, GPT-3.5, Google's Bard (based on PaLM 2), and Anthropic's Claude among the most widely used [41]. In the studies we reviewed, GPT models were predominantly used for extracting clinical information, while BERT models were often employed for protocol-based auto-assessment, though the reasons for these choices were not clearly articulated. Data scarcity continues to pose a significant challenge in utilizing LLMs, potentially affecting the final model performance. To address this issue, data augmentation strategies are crucial, and their implementation varies based on the type, volume, and specific requirements of the data. Similar considerations are applicable when fine-tuning these models. The rationale behind selecting specific strategies for data augmentation and fine-tuning, as well as the data type and use scenarios were transparently reported in the included studies, providing a basis for their transferability to other contexts.

Additionally, while many studies have mentioned the potential of using the LLMs they applied across different contexts, the transferability and scalability of these applications require more standardized reporting. For instance, the questionnaires employed in these studies are varied and inconsistent, frequently self-developed for specific research purposes, which constrains the ability to comprehend the outcomes effectively. Furthermore, most studies predominantly present short-term results, highlighting the need for long-term evaluations, especially in light of the stability issues reported with LLMs. Therefore, there is a need to establish guidelines to evaluate LLMs applications to improve their accuracy and reliability, potentially saving resources before implementation. Transparency in reporting of the development and evaluation of LLMs is essential for their integration into healthcare practices in the future.

Enhancing Interaction and implementation with LLMs

The interaction between users and LLMs in healthcare settings requires careful consideration to ensure successful adoption and user satisfaction [42] [43]. While much of the included research has focused on model performance, the importance of user acceptance and engagement after implementation cannot be overstated. Optimizing the interaction experience is essential for facilitating trust and long-term integration of LLMs into healthcare workflows [44].

Qualitative insights from users regarding AI conversational agents provide invaluable feedback that can guide the effective implementation of these agents in healthcare settings [45]. This is particularly critical in emerging fields, where such insights can help reveal new perspectives. However, our review identified a deficiency in qualitative analysis across most evaluations. Additionally, our findings highlighted a lack of insights into user interactions, which are crucial for refining LLM agents to better meet the behavioral needs of HCPs and patients. As in healthcare contexts, the language used during interactions has been recognized as a key factor influencing user experience and acceptance [46]. Future research could pay greater attention to the qualitative aspects of interacting with LLM agents in managing mental health problems and chronic conditions to inform the interaction design of their integration.

Additionally, understanding how LLM agents interact with existing healthcare systems in implementation process is critical for ensuring their successful integration. Like other AI tools in healthcare, the introduction of LLMs is likely to encounter implementation barriers, such as limitations in contextual understanding, interpretability issues, biases in training data, ethical concerns, the potential impact on HCPs, and regulatory challenges [47]. For instance, Pakhomov et al. highlighted a specific implementation barrier in their development of ANNA, an LLM-based system for intensive monitoring of cognitive changes in cancer patients [40]. While the system aimed to improve care, it also imposed a burden on patients already experiencing significant distress and fatigue due to their treatment, raising concerns about the practicality of such tools in vulnerable populations [40]. But in the included studies we did not find any studies reviewed provide detailed insights into the specific challenges or facilitators encountered during the implementation process, as they are mostly exploration studies.

Lowering the bar for modeling

Same as described in elsewhere [48], LLM could potentially enhance model interpretability by answering individualized questions posed by users. LLMs like ChatGPT, being deep learning models, operate largely as black boxes, making their internal workings difficult to interpret and complicating the assessment of fidelity [49]. However, LLMs have shown potential in demystifying other black-box models for AI models [50]. Our review found that, ChatGPT has shown promise in bridging gaps in programming and data analysis for HCPs when developing deep learning models for diagnosing bone metastasis and analyzing intricate relationships among parameters in diabetic patients [30-38]. We found that LLM agents can help HCPs work across boundaries by lowering the bar for modeling and reducing tedious tasks. The same function were found in studies of using LLM in other conditions as well [51], which help with provide human-understandable insights into the workings of AI models.

Responsible integration of LLMs

Compared to conventional NLP models, the use of LLMs in healthcare has shown promise in addressing a broader range of daily challenges faced by HCPs [45]. However, the inclusion criteria for studies might have limited the scope, as only those with explicit assistive roles were considered. An observation during the screening process is the premature reliance on LLMs for replacing human decision-making processes, such as answering patient inquiries [52-53]. This reliance raises concerns about the accuracy of such replacements, particularly in mental health and chronic conditions due to the lack of validated data in the original databases [54].

Challenges like contextual understanding, interpretability limitations, and biases in training data are well-documented in the use of LLMs in healthcare [47]. Our findings align with these concerns and further emphasize the issue of data scarcity, a recurring theme in several of the included studies. LLMs have been employed to address this challenge by generating synthetic data for data augmentation. This approach to data augmentation has shown significant potential across various medical applications, effectively mitigating the impact of limited data on model performance. By expanding training datasets, LLM-generated data has notably enhanced the accuracy of clinical predictions and diagnoses, improving the robustness and generalizability of models in data-constrained environments.

However, a notable risk emerged as six studies utilized LLMs without any modifications, highlighting the risks of deploying unvalidated AI applications. Inadequate validation can lead to misinformation and AI hallucinations, a limitation noted in several studies [1]. This underscores the importance of continuous monitoring and adjustments to ensure that LLM outputs align with clinical expectations and outcomes. Various mitigation strategies mentioned in the studies include improving

training datasets, simulating adversarial scenarios, enhancing model transparency and interpretability, and incorporating human reviewers into the system [34]. The Human-in-the-Loop (HITL) paradigm offers a collaborative approach where human expertise is integrated with LLM-powered systems, making humans more effective and efficient [55]. While fine-tuning involves adjusting a pre-trained model's parameters to specialize it for a specific task, HITL involves ongoing human involvement in the operation of an AI system, providing guidance, validation, and oversight to improve its performance and adaptability in real-world scenarios.

Future directions

LLM agents are increasingly being leveraged to enhance the quality, affordability, and productivity of chronic condition and mental health services. Their potential is particularly evident in areas that are often overlooked or excluded from routine care, likely due to the limited time and resources available to HCPs to integrate such services into their daily practice [33–36]. Although the development and application of LLMs in chronic condition and mental health services are still in the exploratory phase—evidenced by diverse study designs, evaluation criteria, and a limited number of articles—the prospects for scaling these tools are promising.

Included studies suggested the potential for expanding LLM applications to other conditions, locations, and populations. Specifically, one study proposed scaling up automatic stroke assessment to other domains to reduce physician workload and improve treatment processes [39]. Another suggested integrating LLMs into hospital infrastructures to create scalable mining pipelines for radiologic data [31]. A study highlighted that a clinical physician with rudimentary programming skills could develop deep learning models for medical image analysis for diverse purposes [30]. Furthermore, ChatGPT has demonstrated potential in assisting with the interpretation of scientific data, presenting a valuable resource for researchers with limited computational resources [38].

Limitations

This review is limited by its inclusion of only English-language articles. As LLMs can be tailored to various languages and cultural contexts, our findings may not be applicable to non-English speaking contexts. We also excluded studies that did not explicitly consider the use of LLMs as assistive tools in clinical decision-making, which may have led to the omission of relevant insights and significant applications in this area. Given that the development and application of LLMs in healthcare are still in an exploratory phase, the findings may reflect preliminary and evolving practices. As the field advances, new research may offer different perspectives and more robust evidence. There was notable inconsistency in evaluation standards, dimensions, and methods across the studies reviewed, which makes it challenging to compare findings directly and may affect the overall reliability of the conclusions.

Conclusions

LLM-powered agents hold substantial potential to transform the management of chronic diseases and mental health issues by enhancing health data interpretation, automating clinical assessments, and providing valuable decision-making support. However, the integration of these advanced tools into healthcare systems requires a cautious and responsible approach, given the field's early-stage development and existing challenges such as data limitations and AI hallucinations. A sustainable, long-term adoption will rely on incorporating detailed insights from user interactions and practical implementation experiences.

Future reviews should aim to statistically synthesize findings from various scenarios to establish performance benchmarks for LLMs in managing chronic conditions and mental health issues. Moreover, forthcoming research should take into account the exploration of user interactions, strive

to enhance LLM performance, establish standardized protocols for long-term evaluations, and address existing limitations related to stability and reliability.

Acknowledgements

The authors acknowledge Angela Hardi, a clinical librarian at Becker Medical Library of Washington University School of Medicine, for her assistance in constructing and refining our search strategy. This study was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 956501.

Conflicts of Interest

None declared.

Abbreviations

LLM: Large Language Model

HCP: healthcare professional

AI: Artificial Intelligence

NLP: Natural Language Processing

GPT: Generative Pre-trained Transformer

EHR: Electronic Health Record

References

1. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Communications Medicine* 2023;**3**(1):141 doi: 10.1038/s43856-023-00370-1.
2. Kurniawan MH, Handiyani H, Nuraini T, Hariyati RTS, Sutrisno S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. *Annals of Medicine* 2024;**56**(1):2302980.
3. Sorin V, Glicksberg BS, Artsi Y, et al. Utilizing large language models in breast cancer management: systematic review. *Journal of Cancer Research and Clinical Oncology* 2024;**150**(3):1-8.
4. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large Language Model for Mental Health: A Systematic Review. *arXiv preprint arXiv:2403.15401* 2024.
5. Organization WH. Global status report on noncommunicable diseases; WHO: Geneva, Switzerland, 2014, 2020.
6. Secondary. https://www.who.int/westernpacific/health-topics/noncommunicable-diseases#tab=tab_1.
7. Bloom DE, Cafiero E, Jané-Llopis E, et al. The global economic burden of noncommunicable diseases: Program on the Global Demography of Aging, 2012.
8. The World Bank. Human Development Network. The growing danger of non-communicable diseases. Acting now to reverse course.2011.
9. Picco L, Achilla E, Abdin E, et al. Economic burden of multimorbidity among older adults: impact on healthcare and societal costs. *BMC health services research* 2016;**16**:1-12.
10. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JJY. Randomized Controlled Trials of Artificial Intelligence in Clinical Practice: Systematic Review. *J Med Internet Res* 2022;**24**(8):e37188 doi: 10.2196/37188.
11. Shah-Mohammadi F, Finkelstein J. NLP-Assisted Differential Diagnosis of Chronic Obstructive Pulmonary Disease Exacerbation. *Stud Health Technol Inform*

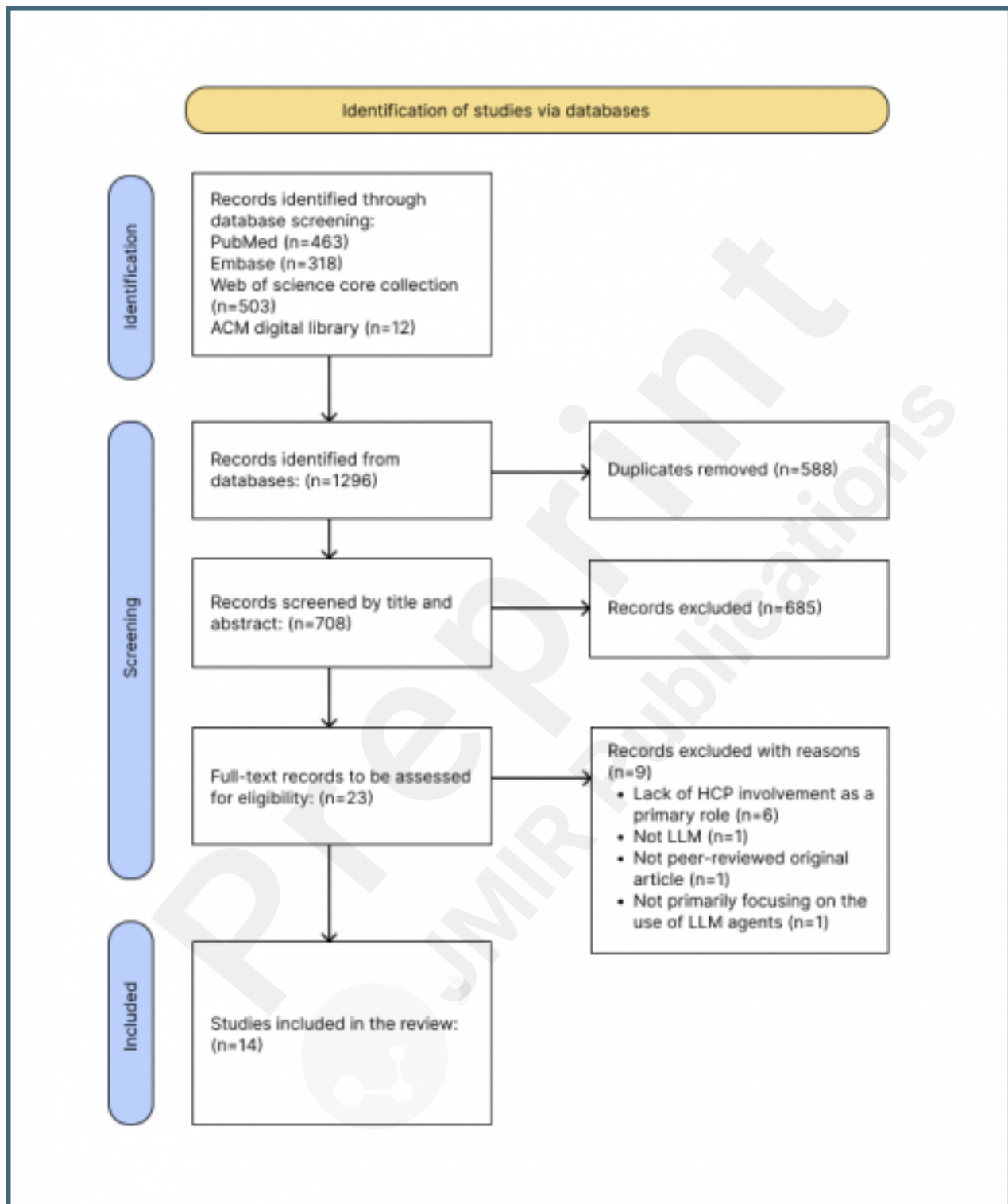
- 2024;**310**:589-93 doi: 10.3233/shti231033.
12. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform* 2019;**7**(2):e12239 doi: 10.2196/12239 [published Online First: 20190427].
 13. Hani SB, Saleh MY. Using Real-Time, Partially Automated Interactive System to Interpret Patient's Data; Helping The Patient To Achieve Diabetic Self-Management: A Rapid Literature Review. *Curr Diabetes Rev* 2023;**19**(5):e311022210519 doi: 10.2174/1573399819666221031161442.
 14. Chapman DP, Perry GS, Strine TW. Peer reviewed: the vital link between chronic disease and depressive disorders. *Preventing chronic disease* 2005;**2**(1).
 15. Le Glaz A, Haralambous Y, Kim-Dufor D-H, et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *J Med Internet Res* 2021;**23**(5):e15708 doi: 10.2196/15708.
 16. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. *Cureus* 2023;**15**(5).
 17. Barlas T, Altinova AE, Akturk M, Toruner FB. Credibility of ChatGPT in the assessment of obesity in type 2 diabetes according to the guidelines. *Int J Obes (Lond)* 2024;**48**(2):271-75 doi: 10.1038/s41366-023-01410-5 [published Online First: 20231111].
 18. Ma Z, Mei Y, Su Z. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. *AMIA Annu Symp Proc* 2023;**2023**:1105-14 [published Online First: 20240111].
 19. Peng W, feng Y, Yao C, et al. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. *Scientific Reports* 2024;**14**(1):2840 doi: 10.1038/s41598-024-52853-3.
 20. Høj S, Thomsen SF, Meteran H, Sigsgaard T, Meteran H. Artificial intelligence and allergic rhinitis: does ChatGPT increase or impair the knowledge? *J Public Health (Oxf)* 2024;**46**(1):123-26 doi: 10.1093/pubmed/fdad219.
 21. Moll M, Heilemann G, Georg D, Kauer-Dorner D, Kuess P. The role of artificial intelligence in informed patient consent for radiotherapy treatments-a case report. *Strahlenther Onkol* 2024 doi: 10.1007/s00066-023-02190-7 [published Online First: 20240105].
 22. Hamed E, Eid A, Alberry M. Exploring ChatGPT's Potential in Facilitating Adaptation of Clinical Guidelines: A Case Study of Diabetic Ketoacidosis Guidelines. *Cureus* 2023;**15**(5):e38784 doi: 10.7759/cureus.38784 [published Online First: 20230509].
 23. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in Senology-An Assessment of Concordance with Breast Cancer Tumor Board Decision Making. *J Pers Med* 2023;**13**(10) doi: 10.3390/jpm13101502 [published Online First: 20231016].
 24. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Medical Education*;n/a(n/a) doi: <https://doi.org/10.1111/medu.15402>.
 25. Arksey H, O'malley L. Scoping studies: towards a methodological framework. *International journal of social research methodology* 2005;**8**(1):19-32.
 26. Levac D, Colquhoun H, O'Brien K. Scoping studies: Advancing the methodology. *Implementation Science* 5: 69, 2010.
 27. Lau C, Zhu X, Chan W-Y. Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry* 2023;**14** doi: 10.3389/fpsy.2023.1160291.
 28. Raile P. The usefulness of ChatGPT for psychotherapists and patients. *Humanities and Social*

- Sciences Communications 2024;**11**(1):47 doi: 10.1057/s41599-023-02567-0.
29. Tan RSYC, Lin Q, Low GH, et al. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. *Journal of the American Medical Informatics Association* 2023;**30**(10):1657-64 doi: 10.1093/jamia/ocad133.
 30. Son HJ, Kim S-J, Pak S, Lee SH. ChatGPT-assisted deep learning for diagnosing bone metastasis in bone scans: Bridging the AI Gap for Clinicians. *Heliyon* 2023;**9**(12):e22409 doi: <https://doi.org/10.1016/j.heliyon.2023.e22409>.
 31. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology* 2023;**308**(3):e231362 doi: 10.1148/radiol.231362.
 32. Lechien JR, Chiesa-Estomba C-M, Baudouin R, Hans S. Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. *European Archives of Oto-Rhino-Laryngology* 2024;**281**(4):2105-14 doi: 10.1007/s00405-023-08326-w.
 33. Wu DJH, Bibault J-E. Pilot applications of GPT-4 in radiation oncology: Summarizing patient symptom intake and targeted chatbot applications. *Radiotherapy and Oncology* 2024;**190**:109978 doi: <https://doi.org/10.1016/j.radonc.2023.109978>.
 34. Gebrael G, Sahu KK, Chigarira B, et al. Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0. *Cancers* 2023;**15**(14):3717.
 35. Choi HS, Song JY, Shin KH, Chang JH, Jang B-S. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J* 2023;**41**(3):209-16 doi: 10.3857/roj.2023.00633.
 36. Lian R, Hsiao V, Hwang J, et al. Predicting health-related quality of life change using natural language processing in thyroid cancer. *Intell Based Med* 2023;**7** doi: 10.1016/j.ibmed.2023.100097 [published Online First: 20230315].
 37. Extracting Thyroid Nodules Characteristics from Ultrasound Reports Using Transformer-based Natural Language Processing Methods. *AMIA Annual Symposium Proceedings*; 2023. American Medical Informatics Association.
 38. Albajy MA, Mernea M, Mihaila A, Pop C-E, Mihăilescu DF. Harnessing Code Interpreters for Enhanced Predictive Modeling: A Case Study on High-Density Lipoprotein Level Estimation in Romanian Diabetic Patients. *Journal of Personalized Medicine* 2023;**13**(10):1466.
 39. Gu Z, He X, Yu P, et al. Automatic quantitative stroke severity assessment based on Chinese clinical named entity recognition with domain-adaptive pre-trained large language model. *Artificial Intelligence in Medicine* 2024;**150**:102822 doi: <https://doi.org/10.1016/j.artmed.2024.102822>.
 40. Pakhomov S, Solinsky J, Michalowski M, Bachanova V. A Conversational Agent for Early Detection of Neurotoxic Effects of Medications through Automated Intensive Observation. *Pac Symp Biocomput* 2024;**29**:24-38.
 41. Naveed H, Khan AU, Qiu S, et al. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* 2023.
 42. Leng L. Challenge, integration, and change: ChatGPT and future anatomical education. *Medical Education Online* 2024;**29**(1):2304973 doi: 10.1080/10872981.2024.2304973.
 43. Shan Y, Ji M, Xie W, et al. Language Use in Conversational Agent-Based Health Communication: Systematic Review. *J Med Internet Res* 2022;**24**(7):e37403 doi: 10.2196/37403.
 44. Xu Y, Gao W, Wang Y, Shan X, Lin Y-S. Enhancing user experience and trust in advanced LLM-

- based conversational agents. *Computing and Artificial Intelligence* 2024;**2**(2):1467-67.
45. Milne-Ives M, de Cock C, Lim E, et al. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *J Med Internet Res* 2020;**22**(10):e20346 doi: 10.2196/20346 [published Online First: 20201022].
 46. Katz NT, Jones J, Mansfield L, Gold M. The Impact of Health Professionals' Language on Patient Experience: A Case Study. *J Patient Exp* 2022;**9**:23743735221092572 doi: 10.1177/23743735221092572 [published Online First: 20220407].
 47. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. *Diagn Pathol* 2024;**19**(1):43 doi: 10.1186/s13000-024-01464-7 [published Online First: 20240227].
 48. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. *Health Care Science* 2023;**2**(4):255-63 doi: <https://doi.org/10.1002/hcs2.61>.
 49. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-Generated Information on Head and Neck and Oromaxillofacial Surgery: A Multicenter Collaborative Analysis. *Otolaryngol Head Neck Surg* 2024;**170**(6):1492-503 doi: 10.1002/ohn.489 [published Online First: 20230818].
 50. Wu X, Zhao H, Zhu Y, et al. Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era. *ArXiv* 2024;**abs/2403.08946**.
 51. Munaf U, Ul-Haque I, Arif TB. ChatGPT: A Helpful Tool for Resident Physicians? *Academic Medicine* 2023;**98**(8).
 52. Yoshiyasu Y, Wu F, Dhanda AK, Gorelik D, Takashima M, Ahmed OG. GPT-4 accuracy and completeness against International Consensus Statement on Allergy and Rhinology: Rhinosinusitis. *Int Forum Allergy Rhinol* 2023;**13**(12):2231-34 doi: 10.1002/alr.23201 [published Online First: 20230616].
 53. Ozgor F, Caglar U, Halis A, et al. Urological Cancers and ChatGPT: Assessing the Quality of Information and Possible Risks for Patients. *Clin Genitourin Cancer* 2024;**22**(2):454-57.e4 doi: 10.1016/j.clgc.2023.12.017 [published Online First: 20240105].
 54. Saeidnia HR, Hashemi Fotami SG, Lund B, Ghiasi N. Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact. *Social Sciences* 2024;**13**(7):381.
 55. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 2023;**56**(4):3005-54.

Supplementary Files

Untitled.



Multimedia Appendixes

PRISMA ScR Checklist.

URL: <http://asset.jmir.pub/assets/7bcd89efa9361f1129def374fbaf490.pdf>



CONSORT (or other) checklists

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
URL: <http://asset.jmir.pub/assets/99858eee53dd53cc6a02d6ab5d662bc6.pdf>