Journal of the American Medical Informatics Association, 2024, **31(8)**, 1743–1753
https://doi.org/10.1093/jamia/ocae145
Advance access publication 20 June 2024
Research and Applications

Research and Applications

Zero-shot learning to extract assessment criteria and medical services from the preventive healthcare guidelines using large language models

Xiao Luo, PhD[1,2,*], Fattah Muhammad Tahabi, MS[1], Tressica Marc, BSc[3], Laura Ann Haunert, MPH[4], Susan Storey, PhD[4]

[1]Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, Stillwater, OK 74078, United States, [2]Department of Biostatistics and Health Data Science, School of Medicine, Indiana University, Indianapolis, IN 46202, United States, [3]Department of Computer Information Technology, Purdue School of Engineering and Technology, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, United States, [4]School of Nursing, Indiana University, Indianapolis, IN 46202, United States

*Corresponding author: Xiao Luo, PhD, Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, 449 Business Building, Stillwater, OK 74078, United States (xiao.luo@okstate.edu)

Abstract

Objectives: The integration of these preventive guidelines with Electronic Health Records (EHRs) systems, coupled with the generation of personalized preventive care recommendations, holds significant potential for improving healthcare outcomes. Our study investigates the feasibility of using Large Language Models (LLMs) to automate the assessment criteria and risk factors from the guidelines for future analysis against medical records in EHR.

Materials and Methods: We annotated the criteria, risk factors, and preventive medical services described in the adult guidelines published by United States Preventive Services Taskforce and evaluated 3 state-of-the-art LLMs on extracting information in these categories from the guidelines automatically.

Results: We included 24 guidelines in this study. The LLMs can automate the extraction of all criteria, risk factors, and medical services from 9 guidelines. All 3 LLMs perform well on extracting information regarding the demographic criteria or risk factors. Some LLMs perform better on extracting the social determinants of health, family history, and preventive counseling services than the others.

Discussion: While LLMs demonstrate the capability to handle lengthy preventive care guidelines, several challenges persist, including constraints related to the maximum length of input tokens and the tendency to generate content rather than adhering strictly to the original input. Moreover, the utilization of LLMs in real-world clinical settings necessitates careful ethical consideration. It is imperative that healthcare professionals meticulously validate the extracted information to mitigate biases, ensure completeness, and maintain accuracy.

Conclusion: We developed a data structure to store the annotated preventive guidelines and make it publicly available. Employing state-of-the-art LLMs to extract preventive care criteria, risk factors, and preventive care services paves the way for the future integration of these guidelines into the EHR.

Key words: preventive care; clinical guidelines; large language models; information extraction.

Introduction

Most health systems globally were designed to be reactive, which means the system was designed to diagnose and treat illnesses rather than prevent the onset of a disease. In the United States, 90% of the nation's $3.3 trillion annual healthcare expenditures are for people with chronic and mental health conditions.[1] Therefore, preventing diseases is key to improving people's health and controlling the rise in health costs. Towards promoting preventive care and clinical services, the United States Preventive Services Task Force (USPSTF)[2] has been established to identify scientific, evidence-based recommendations or guidelines on dozens of clinical preventive services that are intended to reduce the risk for heart disease, cancer, infectious diseases as well as improve the health of children, adolescents, adults, and pregnant women. The integration of these preventive guidelines with Electronic Health Records (EHRs) systems, coupled with the generation of personalized preventive care recommendations, holds significant potential for improving healthcare outcomes. However, there limited research exists on addressing the challenges in automating the processing of these long preventive care guidelines and integrating them into the clinical decision support (CDS) component of EHR systems. Our research endeavors to address this gap by automating the extraction of assessment criteria and risk factors from these guidelines, aligning with the typical EHR data structure. We employed Large Language Models (LLMs) to facilitate this process, enabling the integration of extracted data into the EHR CDS component. Subsequently, these data can be analyzed against patient medical records to generate personalized preventive care recommendations.

Received: March 20, 2024; **Revised:** April 30, 2024; **Editorial Decision:** May 30, 2024; **Accepted:** June 3, 2024

The existing EHR systems have implemented CDS modules to consider the preventive guidelines published by the USPSTF to handle the prevention of major diseases. Clinicians receive reminders when patients are due to take screening tests or exams. The USPSTF guidelines and Healthcare Effectiveness Data and Information Set (HEDIS)[3] technical specifications are updated to the revised evidence-based guidelines and are frequently used to measure provider and health system performance on quality of care. Yet alignment between EHR and quality metrics is often missing.[4] Many clinical criteria for preventive care are specified in the HEDIS technical specifications. However, not all the preventive care criteria are adapted and implemented in EHR decision support systems.[5] On the other side, the criteria in most of the preventive care CDS modules in the EHR are limited to age, gender, and screening intervals that are defined in the USPSTF guidelines. This "one size fits all" preventive CDS does not provide personalized recommendations that take into account risk factors that relate to the patient's family history (eg, family history of diabetes), social history (eg, tobacco use, housing and food insecurity, drug use, alcoholic use, physical exercise, and other social determinants of health [SDOH]), and non-acute disease history. Social history, including behavioral and environmental determinants, are increasingly recognized as key factors for many causes of disease, disability, and mortality in the United States.[6] The WHO states that "The primary prevention services and activities include the provision of information on behavioral and medical health risks and measures to reduce risks at the individual and population levels."[7]

One main challenge to automating the integration of the guidelines into the EHR is information extraction from the guidelines. The existing research on preventive care mainly focuses on standardizing the guidelines to support the screening process. The automation of the extraction of risk factors to enable an easy integration with the EHRs to generate patient-centric, personalized recommendations, and actions has not been investigated. Serban et al[8] developed a pattern extraction approach based on an ontology-driven linguistic pattern identification to reconstruct the knowledge in the guidelines automatically. Peleg and Tu[9] compared 5 different design patterns of clinical guidelines and proposed new design patterns for 2 types of preventive care guidelines, namely, screening guidelines and immunization guidelines. The previous research[10] investigated an Natural Language Processing (NLP) pipeline to automate the information extraction from the preventive care guidelines to enable patient-centric personalized preventive care recommendations. With the advance and research in the LLMs and their application to biomedical text processing, much research has been done to show the superior capability of the LLMs in understanding and processing biomedical text including clinical name entity recognition,[11–13] clinical question answering,[12] and clinical image to text generation,[14] etc. Few research studies focus on information extraction from the clinical guidelines including the preventive care guidelines. Because the number of clinical guidelines for preventive care is limited, there are not enough data instances for model training or fine tuning. So, in this research, we take the zero-shot learning approach for information extraction.

In this article, we report an empirical study of employing the most recent 4 LLMs (GPT3.5,[15] GPT4,[15] and PaLM2[16]) on information extraction from the preventive care guidelines. The objectives were to: (1) annotate the risk factors and medical services in the adult preventive care guidelines in categories A and B published by the USPSTF, (2) develop a data structure to store the annotated information and make it publicly available, (3) evaluate the state-of-the-art LLMs on extraction the preventive care criteria and risk factors using zero-shot learning, and (4) analyze and discuss the potentials, limitations, and ethical considerations of applying the LLMs for clinical guideline processing and information extraction through performance evaluation and analysis.

Methods

Preventive care guideline data

Our study focuses on information extraction from USPSTF guidelines rated as "A" and "B" applicable to adults excluding the pregnancy-related and children-related guidelines. The reason to consider categories "A" and "B" is that the Affordable Care Act (ACA) requires non-grandfathered private health plans to provide coverage without cost-sharing for preventive services rated as "A" (strongly recommended) or "B" (recommended) by the USPSTF.[17,18] Hence, general adults with insurance are eligible to receive these preventive services. Based on the published guidelines by USPSTF by September 2022, we have included 24 guidelines in this research.

Guideline annotation process

To ease the integration of the guideline with the EHR for recommendation generation, we proposed to annotate the information in the guidelines and categorize it into a patient-centered data structure (shown as Figure 1), which includes 5 parts: guideline info, demographics, medical history, SDOH (or social history), family history, and preventive healthcare services. The content in this structure can be compared against the components defined in the schema of Fast Healthcare Interoperability Resources (FHIR).[19]

The demographics include age, gender, ethnicity, or race. Since few guidelines mention ethnicity or race as risk factors, they were grouped into 1 category under demographics. Medical history includes disease or health conditions that increase the risks of developing the disease that is described in the preventive guideline. The SDOH include the descriptions of the conditions that are related to how people are born, live, learn, work, play, and worship that affect a wide range of health, functioning, and quality-of-life outcomes and risks. Family history is the medical or social history of the family members of the patient. Preventive care services include laboratory tests, medical procedures of medication, medical imaging, and counseling or assessment services that a patient needs to engage in order to prevent the development of a disease. The elements of ethnicity or race, medical history, SDOH, and family history are often associated with the increases in the risks of developing a preventable disease. For instance, the lung cancer screening guideline states that "The USPSTF recommends annual screening for lung cancer with low-dose computed tomography (LDCT) in adults aged 55-80 years who have a 30-pack-year smoking history and currently smoke or have quit within the past 15 years." The demographics annotated from this content is age "55-80 years," healthcare service is "LDCT," and "30 pack-year smoking history," "currently smoke," and "quit within the

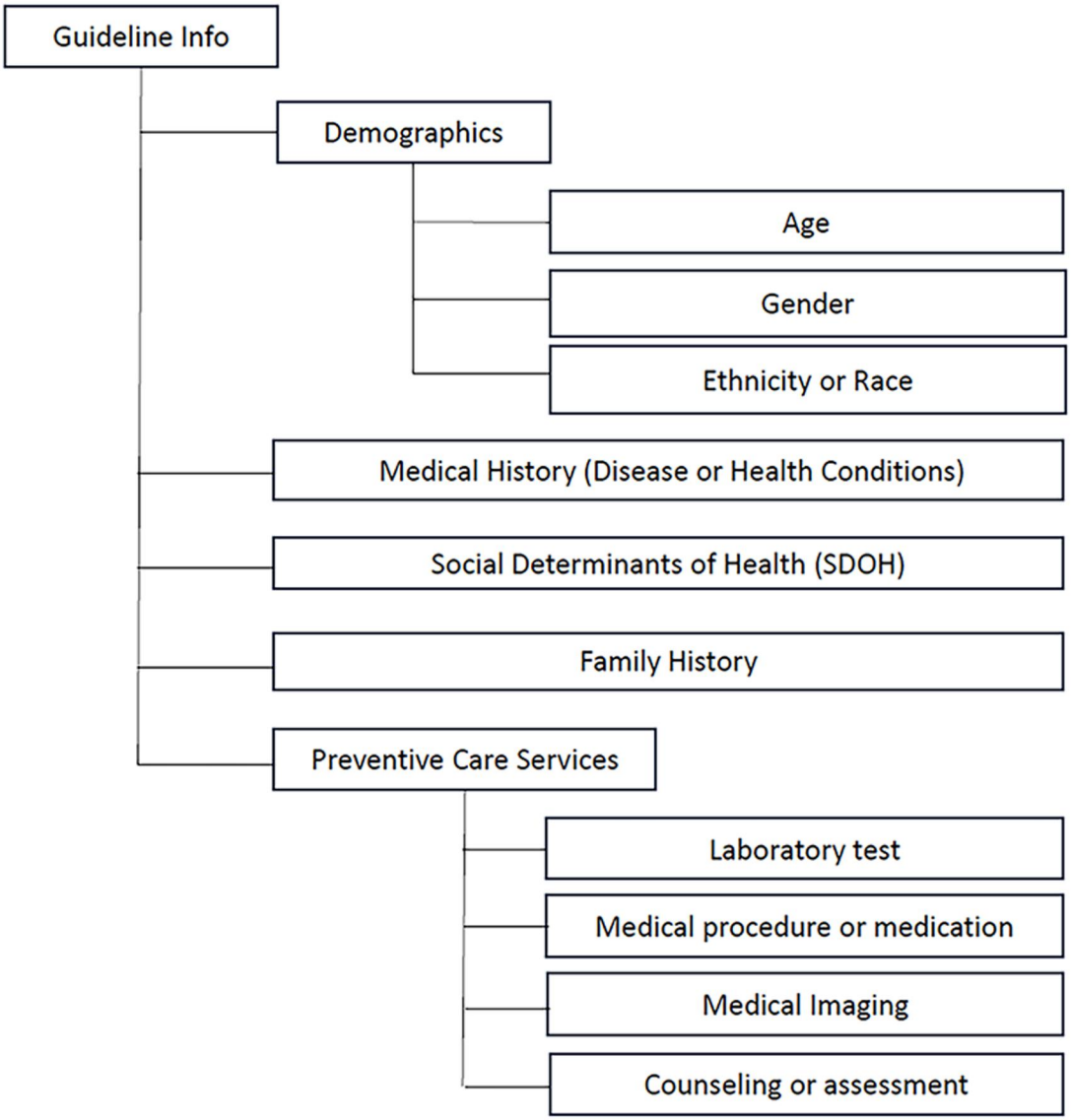

Figure 1. A patient-centered data structure for preventive care guideline annotation.

past 15 years" are SDOH that increase the risks of developing lung cancer. We used a 2-step process to annotate each of the preventive care guidelines. First, one of the authors (S.S.) conducted the initial screening of all guidelines and identified the sections (eg, Population, Risk Assessment parts under the Clinical Summary section, the Patient Population Under Consideration, and Assessment of Risk sections under the Clinical Consideration, etc.) of the guidelines that might: (1) contain clinically relevant information according to the patient-centered data structure, or (2) contain the rationale or the risk assessment (eg, "Black persons and individuals of non-Hispanic ethnicity are at increased risk of anxiety disorders due to social, rather than biological, factors"). Then, 2 annotators (S.S. and L.H.) independently annotated the guidelines and classified the annotated content into the categories described in Figure 1. If the same concept is phrased differently and occurs in different places, they were all annotated as ground truth. For example, both "elevated blood pressure" and "hypertension" were annotated as risk factors in the medical history for the guideline "Healthy Diet and Physical Activity for Cardiovascular Disease Prevention in Adults with Cardiovascular Risk Factors." Inception[20] is used as a tool for data annotation. An example of the annotation is given in Figure S1 in File S2. In this second step, the inter-rater agreement (Cohen's Kappa) is achieved between the 2 coders: 0.84 when determining the annotated content and 0.90 when determining the corresponding category of the content. The disagreed cases were resolved in meetings involving 3 main authors (X.L., S.S., and L.H.). The annotated guidelines are stored in JSON files and included as File S1.

System framework and LLMs for information extraction

Figure 2 shows the system framework and the process of information extraction using the LLMs.

The guideline annotation process is applied to generate a set of JSON files corresponding to the defined data structure (shown as Figure 1). The APIs of the LLMs were used to extract information corresponding to categories "age," "gender," "ethnicity or race," "medical history," "SDOH," "family history," "laboratory test," "medical procedure or

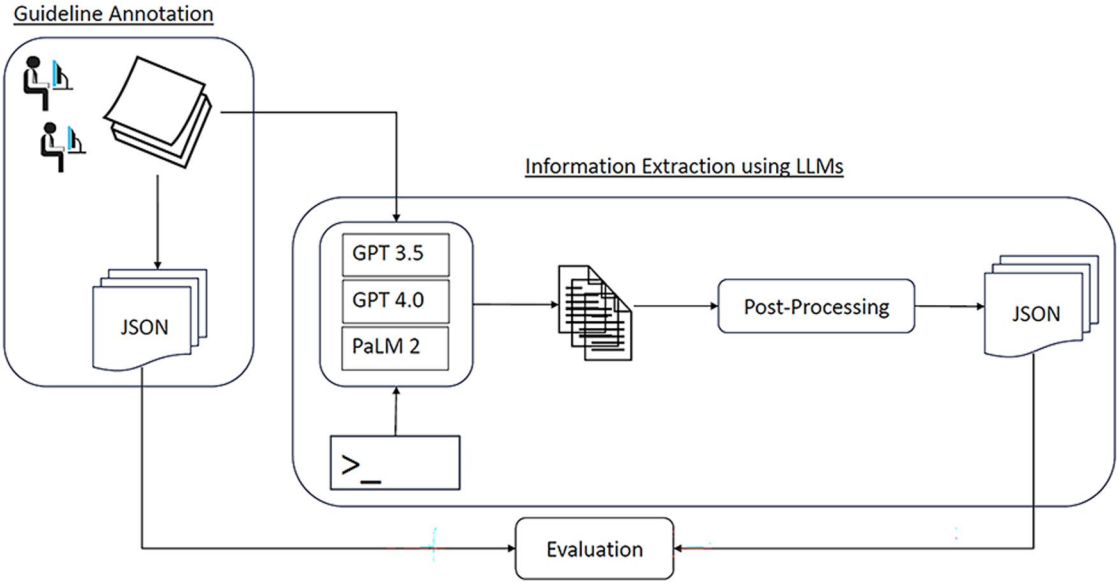

Figure 2. System framework for data annotation, information extraction, and evaluation.

medication," "medical imaging," "counseling or assessment." The API allows users to furnish instructions using 2 role variables. Our prompt is organized intuitively with the following sequence:

- System: outlines task instructions for the LLM in the specified role. We utilized the system variable to supply task instructions, directing the model to assume the role of an annotator.
- User: inputs text for zero-shot learning.

In accordance with the definitions, the user submits a message that incorporates a preventive guideline for annotation and specifies that the assistant should provide a response. The design of the prompt utilized in this research is depicted in Figure 3. The prompt is deliberately uncomplicated, providing minimal information to direct the LLMs in processing the input text and subsequently responding to questions in a predefined format, encompassing the required information for extraction. It is worth noting that the maximum input length in terms of number of tokens of the LLMs were less than the length of the guidelines. Hence, we had to split the guidelines into chunks and input each at a time. The LLMs are instructed to know that input is done when the [END_TOKEN] is given. After finishing the input content, the LLMs are instructed to start answering the questions. The prompt is designed to have LLMs to utilize the original content in the input to produce short answers for each question.

Post-processing

To ensure comparability in evaluation, we implemented a set of standardized postprocessing steps to both ground truth and extracted information to normalize the text representation. First, general text normalization procedures such as lowercasing, whitespace, and hyphens removal, and lemmatization are applied. For categories "age" and "gender," we converted "women" to "female," and "men" to "male," and normalized various age expressions to a format similar to "45-80 years old" or "18 or older." For those terms that are connected via conjunction words, eg, "or," etc., we converted them into a set of terms by removing the conjunction words for comparison between the ground truth and extracted information. For example, the extracted information "screening counseling or assessment interventions" is converted to 2 terms "screening counseling" and "assessment interventions" for evaluation. Considering that abbreviations are often used in the medical context, we also generated abbreviations for the ground truth and extracted terms for comparison. For example, "STI" is generated from "sexually transmitted infection" before evaluation.

Performance evaluation metrics

To evaluate the performance, the evaluation metrics that are applied to the Named Entity Recognition (NER) tasks were used. Based on previous research,[21] in the scenario of zero-shot learning, it is also informative to use semantic evaluation. For example, if the annotated ground truth in the original text is "African American," and the extracted concept by the LLMs is "Black American," we consider the extraction to be semantically correct. Therefore, in this study, we utilize both NER evaluation metrics (Recall [R], Precision [P], and F1) and a semantic evaluation approach. In the NER evaluation metrics, we implemented a relaxed match, akin to the fragment match employed in related research on biomedical NER.[11,22,23] The equations of the calculations are given in the File S2.

In the semantic evaluation, we initially transform both the annotated ground truth and the extracted concept into embeddings using the sentence-transformers model (specifically, all-MiniLM-L6-v22124). Subsequently, we gauge the cosine similarity between these embeddings. If the cosine similarity (s) surpasses a designated threshold θ (eg, $\theta = 0.8$ or 0.9), it is deemed a match; otherwise, it is categorized as an incorrect extraction. We compute the average cosine similarity (Ave [s]) for each type without introducing a threshold and determine accuracy using the specified threshold θ, set at 0.8.

In addition to the evaluation of information extraction from guidelines, we also evaluated the overall performance of accuracy at the guideline level. A guideline often does not

Role	Content
System	After all content of a document is given, please answer a set of questions based on the given document. Since the document is very long, the input token [END_TOKEN] indicate the end of the document. The answers to the questions should be returned as a list in format [A1:, A2:, A3:, ...]. Please only use the original words from the input document. Each answer be consisted of a set of phrases. An output example: [A1: 40 years and older, A2: female, ...]
User	*input narrative* [END_TOKEN] Question 1: What is the age group of patients to recommend for Abdominal Aortic Aneurysm Screening? Question 2: Which gender should be recommended for Abdominal Aortic Aneurysm Screening? ..... Question n: Which laboratory test or medical exam should be recommended for Abdominal Aortic Aneurysm Screening?

Figure 3. Prompt to instruct LLMs to extract information from the guidelines.

contain criteria or information for all categories. Given a guideline, if there is a total of n categories, and information from m categories is correctly extracted, the accuracy is calculated using eqn (1).

$$\text{Accuracy} = \frac{m}{n}. \quad (1)$$

Results

Descriptive statistics of the annotated guidelines

We annotated 24 preventive care guidelines. Table 1 shows the statistics of the guidelines. 78.26% of the guidelines ($n = 18$) are applicable to both males and females. When no specific gender is mentioned in the guideline, it is applicable to both genders. 39.13% of the guidelines are applicable to the population older than 35 years. 13.04% of the guidelines mention certain races or ethnicities have a higher risk of those preventable diseases. Only one guideline—"Weight Loss to Prevent Obesity-Related Morbidity and Mortality in Adults" applies to adults with a body mass index (BMI) of 30 or higher. 69.57% of the guidelines are associated with some diseases related factors, such as cardiovascular disease, or diabetes, etc. 65.22% of the guidelines have described the associated risk factors of SDOH. 39.13% of the guidelines have described the associated risk factors of family history. For example, the colorectal cancer screening guideline specified that the medical history of obesity and diabetes, SDOH of long-term smoking and unhealthy alcohol use, and the family history of colorectal cancer are the risk factors of colorectal cancer. The preventive care services include various laboratory tests, medical procedures or medication, medical imaging, counseling, or a combination of them. For example, the guideline "Tobacco smoking cessation in adults" lists using a combination of behavioral counseling sessions and nicotine replacement therapy.

NLP statistics of the information in each category

With the objective to automate the information extraction from the guidelines based on the criteria and risk factors, we analyzed the complexity of the information in these categories which might impact the performance of the LLMs. Table 2 shows the NLP statistics of the total number of terms in the ground truth in all guidelines and average length in terms of the number of words in each category. The SDOH, family history, laboratory test, and counseling have more words in the description. Whereas gender and ethnicity or race have fewer number of words in representation. Most age criteria have a representation of 3 words with one exception that has more than 3 words—"25 or older at risk."

Table 1. Statistics of the guidelines based on the criteria.

Categories and values	Number of guidelines
Gender	
Male	1 (4.17%)
Female	4 (16.67%)
Both	19 (79.17%)
Age groups	
18 or older or 18-79	12 (50%)
21 or older	3 (12.5%)
35 or older	2 (8.33%)
40 or older	1 (4.17%)
45 or older	1 (4.17%)
50 or older	2 (8.33%)
65 or older	3 (12.5%)
Ethnicity or race (risk factor)	3 (12.5%)
Medical history	
Required condition	1 (4.17%)
Risk factor	16 (66.67%)
SDOH	15 (62.5%)
Family history	9 (37.5%)
Preventive care services	
Laboratory test	15 (62.5%)
Medical procedure or medication	6 (25%)
Medical imaging	2 (8.33%)
Counseling	9 (37.5%)

Performance comparison on information extraction

Table 3 shows the performances of GPT3.5, GPT4, and PaLM2 models using the NER evaluation metrics. Each of the LLM shows its advantage in extracting information for some categories. GPT3.5 works better than the other 2 on categories of ethnicity or race, SDOH, and medical imaging, whereas GPT4 works better on categories of age, gender, family history, medical procedure or medication, and counseling. PaLM2 works better in the categories of medical history and laboratory tests. GPT3.5 has a slightly higher average performance overall. The GPT 3.5 can extract most of the mentioned ethnicity or race information correctly, except "American Indian/Alaskan Native" which was mentioned in the guideline of "Screening for Colorectal Cancer."

Table 2. Summary of the criteria of the guidelines.

Category	Number of terms	Average length	Examples
Age	23	3.087	"18 or older," "40-74"
Gender	5	1.000	"female," "male"
Ethnicity or race	4	1.750	"Hispanic," "African Americans," "American Indians"
Medical history	38	2.079	"chronic illnesses," "mental health disorders"
SDOH	36	3.361	"long-term smoking," "unhealthy alcohol use"
Family history	12	3.083	"breast cancer," "ovarian cancer"
Laboratory test	25	3.960	"fasting plasma glucose," "HPV testing"
Medical procedure or medication	14	2.643	"tamoxifen," "nicotine replacement therapy"
Medical imaging	2	3.500	"low-dose computed tomography"
Counseling	27	3.333	"genetic counseling," "assessments of gait and mobility"

Table 3. Performance comparison using NER evaluation metrics.

	GPT 3.5[a]			GPT 4[a]			PaLM 2[a]		
Category	*R*	*P*	F1	*R*	*P*	F1	*R*	*P*	F1
Age	0.507	.507	0.507	**0.787**	**.787**	**0.787**	0.606	.503	0.538
Gender	1.000	.866	0.900	**1.000**	**1.000**	**1.000**	1.000	.732	0.800
Ethnicity or race	**0.917**	**.917**	**0.917**	0.333	.083	0.133	0.583	.583	0.583
Medical history	0.255	.218	0.227	0.375	.428	0.376	**0.416**	**.402**	**0.386**
SDOH	**0.461**	**.494**	**0.471**	0.284	.249	0.257	0.433	.447	0.418
Family history	0.561	.556	0.547	**0.630**	**.601**	**0.580**	0.583	.474	0.510
Laboratory test	0.444	.447	0.427	0.471	.481	0.467	**0.606**	**.565**	**0.548**
Medical procedure or medication	0.390	.390	0.390	**0.593**	**.593**	**0.593**	0.318	.252	0.277
Medical imaging	**1.000**	**1.000**	**1.000**	0.750	.750	0.750	1.000	.700	0.795
Counseling	0.363	.283	0.302	**0.406**	**.450**	**0.400**	0.392	.388	0.377
Macro-average	**0.590**	**.568**	**0.569**	0.563	.542	0.534	0.594	.505	0.523

[a] The bold numbers indicate the best-performing model in each corresponding category.

Table 4. Performance comparison using semantic similarity-based accuracy (SSA).

Category	GPT 3.5[a]	GPT 4[a]	PaLM 2[a]
Age	0.591	**0.773**	0.591
Gender	0.800	**1.000**	0.600
Ethnicity or race	**0.750**	0.250	0.667
Medical history	0.588	0.529	**0.706**
SDOH	0.800	0.400	**0.867**
Family history	0.444	**0.667**	0.556
Laboratory test	0.600	0.600	**0.800**
Medical procedure or medication	0.500	**0.667**	**0.667**
Medical imaging	**1.000**	0.500	**1.000**
Counseling	0.444	0.444	**0.778**
Macro-average	0.677	0.558	**0.723**

[a] The bold number indicates the best-performing model in each corresponding category.

The GPT4 extracted "Black individuals" instead of "African American" and extracted "Hispanic women" instead of "Hispanic," which lowered the performances of GPT4.

Table 4 shows the performance comparison when semantic similarity evaluation is applied. We calculated the accuracy by setting the cosine similarity threshold to 0.8, which means when the cosine similarity between the ground truth and extracted terms is greater than 0.8, it is a correct match. The best overall performance in terms of accuracy (0.723) is gained by PaLM2. By using this evaluation metric, the performance of the LLMs in some categories increased. For example, the performance on SDOH increased to over 0.8 when GPT3.5 or PaLM2 was used, and the performance on counseling service increased to 0.778 when PaLM2 was used. The reason is that semantic-based evaluation can capture those concept representations with similar or same semantic meaning but using different words. For example, the SDOH of the guideline "Hepatitis C Virus Infection in Adolescents and Adults Screening" includes "born to an HCV-infected mother," but the extracted information by PaLM2 is "having a mother who had HCV during pregnancy." The cosine similarity between these 2 terms is 0.859, whereas the NER based F1 is low.

Performance comparison on complete information extraction on guidelines

Figure 4 shows the performances of the LLMs on extraction information for the guidelines when semantic similarity evaluation metric is used. Each guideline has a set of criteria, risk factors, and preventive care services to be extracted. The *x*-axis shows the guidelines with the number of annotated categories (defined in Figure 1) to be extracted. The guideline "Osteoporosis to Prevent Fractures: Screening" has as many as 7 categories of information to be extracted, which include age, gender, medical history, family history, SDOH, and 2 types of preventive care exam or service. Whereas the guideline "Tobacco Smoking Cessation in Adults: Screening" only has 2 categories of information to be extracted, which includes age and preventive counseling service. For some guidelines, the majority of the LLMs were able to extract information from all categories to make the accuracy as high as 1. For example, all 3 LLMs can extract all 4 categories of information for Breast Cancer Screening guideline, and GPT3.5 and PaLM2 can extract all categories of information for "Hepatitis C Virus Infection screening" and "Hypertension Screening" guidelines. For the guideline

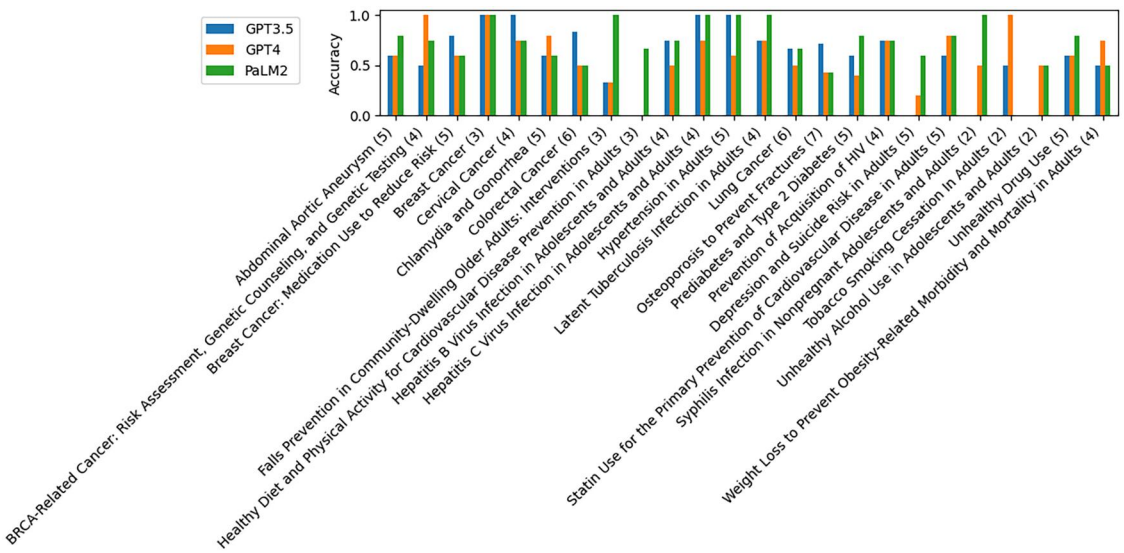

Figure 4. Performance comparison on guidelines.

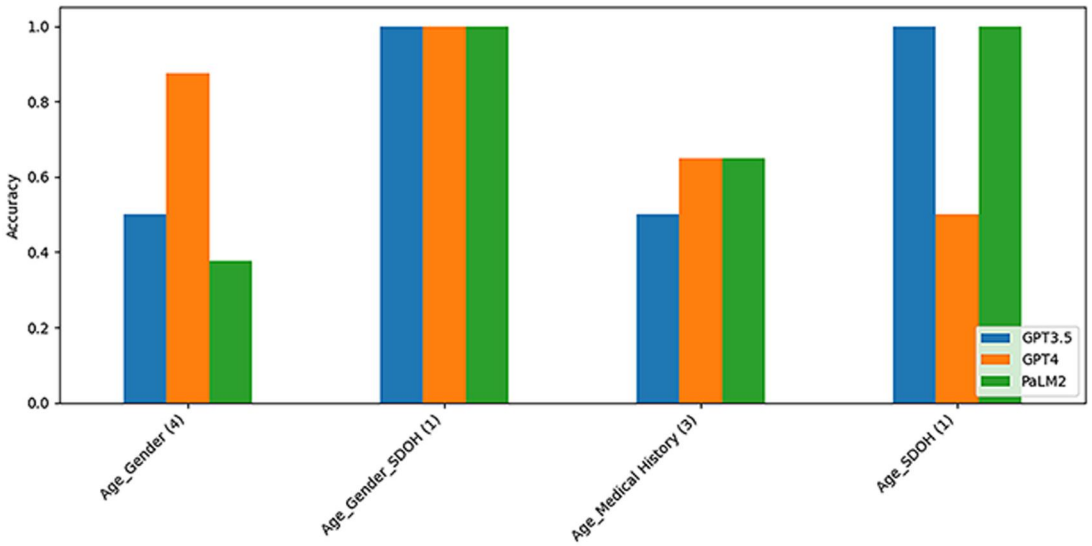

Figure 5. Performance on guidelines with connected assessment criteria.

"Unhealthy Alcohol Use in Adults Screening," 2 of the LLMs—GPT4 and PaLM2 could extract the age group correctly. None of the LLMs were able to extract the specific recommended unhealthy alcohol use counseling services, such as Tweak alcohol screening test,[25] T-ACE screening tool,[26] etc. Comparing the 3 LLMs, neither of them works significantly better than the others in terms of overall accuracy based on the semantic similarity evaluation metric.

It is worth noting that there are guidelines specifying that the patients need to meet 2 or more assessment criteria for screening or preventive care. Upon analysis, we discovered that among the 24 guidelines, 9 contain interconnected assessment criteria spanning 2 or 3 categories. Gender and age are the most frequently connected assessment criteria. The only guideline with 3 connected assessment criteria—age, gender, and SDOH—is "Abdominal Aortic Aneurysm Screening." Additionally, there are 3 guidelines that have age and medical history (such as cardiovascular disease or diabetes risk factors) as connected assessment criteria. Figure 5 shows the average performance on the number of guidelines with different connected criteria. The number of guidelines in each category is indicated in parentheses on the *x*-axis labels. The performance of the LLMs varies. The lower performance observed on guidelines with age and medical history as criteria is primarily attributed to incomplete extraction of medical history. The performance of GPT3.5 and PaLM2 in accurately extracting the appropriate age range for certain guidelines resulted in lower performance for guidelines having both age and gender as interconnected assessment criteria. For example, GPT3.5 extracted "postmenopausal women" instead of "65 or older" as age group from guideline "Osteoporosis to Prevent Fractures: Screening."

Discussion

LLMs for preventive care guideline processing

Our work is the first to show the LLMs are capable of extracting information from the long clinical guidelines via

zero-shot learning. In contrast to the conventional method of extracting information from these guidelines through human annotation, the LLM-based approach offers a substantial improvement in efficiency. For a single guideline, the average processing time for GPT3.5, GPT4, and PaLM2 are 6.96, 20.51, and 21.99 seconds, respectively. These computations were performed on a machine with a 2-core CPU running at a frequency of 2200 MHz and 12GB of RAM. In contrast, human annotation usually takes around 45 minutes per guideline. It's important to acknowledge that human annotation serves as a reliable ground truth, while information extracted using LLM-based methods may introduce biases, inconsistencies, and therefore requires human validation. The performance of the LLMs on extraction of the basic demographic criteria or risk factors including gender, age range, and ethnicity or race is relatively high, whereas the performance of the LLMs on extraction of the risk factors in category medical history, SDOH, and family history is lower. The performances on extracting the preventive care services in the categories medical imaging and laboratory test are higher than those in the categories of medical procedure or counseling services. We think the main reason is that those medical imaging services and laboratory tests are the most standard and specific ones, whereas some of the medical procedures and counseling services include various non-specific ones. For example, the medical management or interventions that are recommended for falls prevention in community-dwelling older adults include adaptation and modification of home environment, management of postural hypotension, etc. For these situations, the LLMs output a summary intervention as multifactorial interventions or multiple specific ones that might not exactly match those mentioned in the guideline. Another example is the extraction information from the guideline "Healthy Diet and Physical Activity for Cardiovascular Disease Prevention in Adults with Cardiovascular Risk Factors," GPT models struggle primarily due to inaccuracies in extracting the exact age range, medical history risk factors, and counseling services. This includes errors such as setting an upper limit to the age range or failing to extract specific conditions like hypertension or metabolic syndrome, and all counseling services like dietary counseling, physical activity counseling, and behavioral counseling. From the guideline point of view, at least 1 of the LLMs can automate the extraction of all relevant information from 37.5% ($n=9$) of the studied guidelines including breast cancer screening, cervical cancer screening, BRCA-related cancer screening, etc. To investigate whether the integration of the LLMs can improve the overall performance, we took a majority vote approach to merge the output of the LLMs then compared against the ground truth. The results (included in File S2) show that integration via majority vote approach does not make significant improvements. Merging the output of various LLMs via semantic analysis should be considered in future research.

Capability of LLMs on processing long clinical guidelines

When the LLMs were implemented, the length limit in terms of tokens of GPT3.5, GPT4, and PaLM2 were 4096, 8192, and 8K, respectively. Such input limit posed a challenge in adapting the LLMs for processing long clinical guidelines which often include more than 8K tokens. Because of the long input, we had to split the input into chunks for LLMs to analyze. This might break the dependency of the context. On the other side, recent research shows that the limit of the input can introduce memory loss of some LLMs, so that the previous input or interaction with the LLMs were not captured or utilized in producing the output.[27–29] In our study, as suggested by the literature,[30–32] we truncated the input to fit to the size of the token limit and instructed the LLMs to produce results till all input chunks were given. However, we found that sometime the LLMs might not utilize all the input chunks to extract or summarize the correct information. For example, even the implementation section of the colorectal cancer screening guideline states that "Screen all adults aged 45-75 years for colorectal cancer," the GPT3.5 and PaLM2 extracted age range as "50-75." One reason could be that the last input chunk of colorectal cancer screening guideline mentions that "In the current recommendation, while continuing to recommend colorectal cancer screening in adults aged 50-75 years (A recommendation), the USPSTF now recommends offering screening starting at age 45 years (B recommendation)." This might generate challenges for the LLMs to comprehend based on the whole content. Some recent research investigated upgraded LLMs that can analyze the long input for content comprehension, such as LongLM,[33] LongRoPE,[34] etc. We believe these upgraded LLMs can be finetuned towards medical domain and be utilized to extract information from the clinical guidelines.

Generative content and self-consistency of the LLMs

Our objective is to utilize LLMs to extract information from the long preventive care guidelines. Despite attempting various instructions to guide LLMs in responding to questions using the original content present in the provided input text, the LLMs frequently generate content that deviates from the original words in the input. In certain instances, the generative content holds the same meaning or produces an abbreviation of the original input. However, there are occurrences where the generative content cannot be located in the input guidelines. For instance, when examining the guideline "Screening for Hypertension in Adults," which asserts that overweight and obesity are risk factors for developing hypertension, both GPT3.5 and PaLM2 extracted diabetes as disease a risk factor. Although diabetes-related hypertension is a disease mentioned in the literature,[35,36] diabetes was not mentioned in the screening guideline. The primary reason is that the screening guideline is developed for adults 18 years and older without known hypertension. Additionally, similar to recent research, our study reveals a notable issue of low self-consistency in LLMs.[21,37,38] Even if the same instruction and prompt were given, the LLMs can produce different output. The issue of self-consistency raises concerns about the reliability of using LLMs to process lengthy guidelines, emphasizing the importance of thorough evaluation and validation of LLM outputs before implementing them in real-world clinical platforms for physicians or patients. Minor inconsistencies can undermine the trustworthiness of employing LLMs and other AI techniques in healthcare settings. Therefore, addressing the issue of low self-consistency is necessary, a few approaches can be applied: (1) Instruction tuning: Some work in the literature shows that it is necessary to employ instruction fine-tuning to help the LLMs follow the instructions strictly. The instruction fine-tuning can be done using a reward fine-tuning method[39] or ranking the preferred answers;[40] (2) Prompting: Recent research suggests including

a few examples in the prompt or use the most optimal prompt to produce the results multiple times and selecting the most frequent one;[41,42] (3) Training via consistency alignment: A most recent publication proposes initially fine-tuning LLMs with instruction augmentation to enhance the model's ability to generalize following instructions. Subsequently, the model is trained to identify responses that better align with human expectations, utilizing consistency rewards to distinguish the generated responses effectively.[43] Evaluation of these approaches towards information extraction from the guidelines needs to be done in the future.

Practical implication of integrating preventive care guidelines into the EHR

The ultimate goal of this research is to seamlessly integrate the preventive care guidelines into the EHRs system. Our study demonstrates the feasibility of automating the periodic retrieval of guidelines from the USPSTF website through web scraping, followed by the extraction of assessment criteria and risk factors using LLMs. For instance, the "breast cancer screening guideline" can be scraped annually and processed by the LLMs to extract assessment criteria and risk factors based on the predefined prompt. This prompt can be refined to prompt LLMs to provide original guideline content as supporting evidence for each extracted criterion and risk factor, streamlining the human validation process. The extracted data can be organized in the data structure illustrated in Figure 1 to enhance interoperability with the EHR. Following human rigorous validation, the structured data pertaining to breast cancer screening can be seamlessly integrated into the EHR and analyzed alongside patient data to formulate personalized preventive care recommendations. Ensuring the healthcare professionals rigorously validate the extracted criteria and risk factors is crucial to confirming their completeness, accuracy, and consistency. The formulated personalized preventive care recommendations should also be validated before presenting to the patients to maintain accuracy in the communication of medical information.

Ethical considerations of using LLMs in processing healthcare guidelines

As highlighted in recent literature, the integration of LLMs in healthcare decision-making necessitates careful consideration of ethical implications.[44–47] In the context of extracting information from healthcare guidelines for analysis against patient data in the EHR, ethnic considerations converge with ethical concerns related to algorithms, information, and privacy. Given that LLMs are pre-trained using extensive internet datasets, there are apprehensions regarding biases and fairness inherent in the training data.[48,49] As discussed earlier, LLMs may generate information not originally present in the guidelines, potentially introducing biases into the extraction process. For instance, certain racial risk factors may not be adequately captured. Transparency is also a key concern. Given that the healthcare guidelines are often long text documents, it would be beneficial for LLMs to provide supporting evidence from the input guideline that contains the extracted information. So that health professionals can validate the extraction. This also provides transparency and interpretability of the LLMs. Since the goal of extracting information from the preventive care guidelines is to enable personalized and efficient preventive care decision-making, the extracted data must be validated rigorously and oversight by healthcare professionals[50,51] and implemented correctly by system engineers[52] into the EHR system. The generated personalized preventive care recommendations should also be validated by healthcare professionals to confirm the accuracy and protection of information privacy before presenting to patients. The overall framework of extracting information from guidelines and integration into the EHR will need to go through an ethical evaluation and meet the regulatory requirements before the implementation into real-world production.

Limitations and future work

Although this work only evaluated 3 state-of-the-art LLMs towards information extraction from the clinical guidelines, it provides the first insight and foundation on how LLMs can be utilized in processing long clinical guidelines. With the advancing of LLMs, more sophisticated LLMs and fine-tuned LLMs, such as BioMedGPT-LM,[53] PMC-LLaMA,[54] Me LlaMA,[55] etc., that are geared toward clinical domains can be evaluated in future research. Our goal is to automate the integration of the preventive care guidelines with the EHR to generate personalized preventive care recommendations. This work only focused on information extraction from the guidelines, based on the designed data structure. Comparison of the extracted information against the data standard in the EHR can be further evaluated to demonstrate whether generative AI and LLMs can assist in data match and following preventive care recommendations. We focused on zero-shot learning in this work. Although there are a limited number of guidelines, few-shot learning can be implemented in the future to demonstrate whether performance can be improved.

Conclusion

In summary, our research shows that LLMs can be applied to automate the extraction of the criteria and risk factors from the preventive care guidelines. The extracted risk factors in the categories of medical history, SDOH, and family history can be used to analyze and compare against the patient data stored in the EHR systems to generate personalized preventive care recommendations. Personalized disease prevention can improve healthcare outcomes and reduce healthcare expenditures in the long term. Nevertheless, the utilization of LLMs for processing healthcare guidelines for personalized care recommendation generation requires careful ethical consideration. Ensuring that information extraction is validated by healthcare professionals before analysis against patient records is essential to mitigate biases, ensure completeness, and maintain accuracy.

Author contributions

Xiao Luo and Susan Storey conceptualized and designed the study and secured the funding. Xiao Luo, Fattah M. Tahabi, and Susan Storey ensured the integrity and accuracy of the data analysis. Data annotation was carried out by Laura A. Haunert, Susan Storey, and Xiao Luo, whereas Xiao Luo and Tressica Marc handled data cleaning and export to JSON format. Xiao Luo and Fattah M. Tahabi were responsible for model implementation and analysis. Xiao Luo, Fattah M. Tahabi, Laura A. Haunert, and Susan Storey participated in the interpretation and discussion of the findings. Xiao Luo,

Fattah M. Tahabi, and Susan Storey completed the initial draft, and subsequent versions were reviewed and approved by all authors.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the National Institute of General Medical Sciences (NIGMS) (grant number R15GM139094).

Conflicts of interest

None declared.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

References

1. Health and Economic Costs of Chronic Diseases. Accessed October 29, 2023. https://www.cdc.gov/chronicdisease/about/costs/index.htm
2. U.S. Preventive Services Task Force. Accessed October 29, 2023. https://www.uspreventiveservicestaskforce.org/uspstf/
3. Healthcare Effectiveness Data and Information Set (HEDIS) Technical Resources. Accessed October 29, 2023. https://www.ncqa.org/hedis/measures/
4. Bundy DG, Persing NM, Solomon BS, et al. Improving immunization delivery using an electronic health record: the ImmProve project. *Acad Pediatr*. 2013;13(5):458-465.
5. Hatch BA, Tillotson CJ, Huguet N, Hoopes MJ, Marino M, DeVoe JE. Use of a preventive index to examine clinic-level factors associated with delivery of preventive care. *Am J Prev Med*. 2019;57(2):241-249.
6. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. *JAMA*. 2004;291(10):1238-1245. https://doi.org/10.1001/jama.291.10.1238
7. Health promotion and disease prevention through population-based interventions, including action to address social determinants and health inequity. Accessed October 29, 2023. https://www.emro.who.int/about-who/public-health-functions/health-promotion-disease-prevention.html
8. Serban R, ten Teije A, van Harmelen F, Marcos M, Polo-Conde C. Extraction and use of linguistic patterns for modelling medical guidelines. *Artif Intell Med*. 2007;39(2):137-149.
9. Peleg M, Tu SW. Design patterns for clinical guidelines. *Artif Intell Med*. 2009;47(1):1-24.
10. Shah S, Luo X. Extracting modifiable risk factors from narrative preventive healthcare guidelines for EHR integration. In: *Proceedings of 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE; 2017:514-519.
11. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. 2024:ocad259. https://doi.org/10.1093/jamia/ocad259
12. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv, arXiv:230313375, 2023, preprint: not peer reviewed. Published online.
13. Ramachandran GK, Fu Y, Han B, et al. Prompt-based extraction of social determinants of health using few-shot learning. In: *Proceedings of the 5th Clinical Natural Language Processing Workshop*, Toronto, Canada. Association for Computational Linguistics; 2023:385-393.
14. Selivanov A, Rogov OY, Chesakov D, Shelmanov A, Fedulova I, Dylov DV. Medical image captioning via generative pretrained transformers. *Sci Rep*. 2023;13(1):4171.
15. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv, arXiv:230308774, 2023, preprint: not peer reviewed. Published online.
16. Anil R, Dai AM, Firat O, et al. PaLM 2 technical report. arXiv, arXiv:230510403, 2023, preprint: not peer reviewed. Published online.
17. Han X, Robin Yabroff K, Guy GP, Zheng Z, Jemal A. Has recommended preventive service use increased after elimination of cost-sharing as part of the affordable care act in the United States? *Prev Med*. 2015;78:85-91. https://doi.org/10.1016/j.ypmed.2015.07.012
18. Fox JB, Shaw FE. Clinical preventive services coverage and the affordable care act. *Am J Public Health*. 2015;105(1):e7-e10. https://doi.org/10.2105/AJPH.2014.302289
19. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The fast health interoperability resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. *JMIR Med Inform*. 2021;9(7):e21929.
20. De Castilho RE, Klie JC, Kumar N, Boullosa B, Gurevych I. Linking text and knowledge using the inception annotation platform. In: *IEEE 14th International Conference on e-Science*. IEEE; 2018:327-328.
21. Bhate NJ, Mittal A, He Z, Luo X. Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using GPT model. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE; 2023:1476-1480.
22. Luo X, Gandhi P, Storey S, Huang K. A deep language model for symptom extraction from clinical text and its application to extract COVID-19 symptoms from social media. *IEEE J Biomed Health Inform*. 2021;26(4):1737-1748.
23. Chen A, Yu Z, Yang X, Guo Y, Bian J, Wu Y. Contextualized medication information extraction using transformer-based deep learning architectures. *J Biomed Inform*. 2023;142:104370. https://doi.org/10.1016/j.jbi.2023.104370
24. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics; 2019:3982-3992.
25. Cherpitel CJ. Screening for alcohol problems in the US general population: a comparison of the CAGE and TWEAK by gender, ethnicity, and services utilization. *J Stud Alcohol*. 1999;60(5):705-711.
26. Bradley KA, Boyd-Wickizer J, Powell SH, Burman ML. Alcohol screening questionnaires in women: a critical review. *JAMA*. 1998;280(2):166-171.
27. Wang B, Liang X, Yang J, et al. Enhancing large language model with self-controlled memory framework. arXiv, arXiv-2304, 2023, preprint: not peer reviewed. Published online.
28. Packer C, Fang V, Patil SG, Lin K, Wooders S, Gonzalez JE. MemGPT: towards LLMs as operating systems. arXiv, arXiv:231008560, 2023, preprint: not peer reviewed. Published online.
29. Wang X, Salmani M, Omidi P, Ren X, Rezagholizadeh M, Eshaghi A. Beyond the limits: a survey of techniques to extend the context length in large language models. arXiv, arXiv:240202244, 2024, preprint: not peer reviewed. Published online.
30. Xu P, Ping W, Wu X, et al. Retrieval meets long context large language models. In: *The Twelfth International Conference on Learning Representations*. 2024.

31. Laskar MT, Fu XY, Chen C, Tn SB. Building real-world meeting summarization systems using large language models: a practical perspective. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 2023:343-352.
32. Dong Z, Tang T, Li L, Zhao WX. A survey on long text modeling with transformers. arXiv, arXiv:230214502, 2023, preprint: not peer reviewed. Published online.
33. Jin H, Han X, Yang J, et al. LLM maybe longLM: self-extend LLM context window without tuning. arXiv, arXiv:240101325, 2024, preprint: not peer reviewed. Published online.
34. Ding Y, Zhang LL, Zhang C, et al. LongRoPE: extending LLM context window beyond 2 million tokens. arXiv, arXiv:240213753, 2024, preprint: not peer reviewed. Published online.
35. Jia G, Sowers JR. Hypertension in diabetes: an update of basic mechanisms and clinical disease. *Hypertension*. 2021;78 (5):1197-1205.
36. Naha S, Gardner MJ, Khangura D, Kurukulasuriya LR, Sowers JR. Hypertension in diabetes. In: *Endotext*. MDText.com, Inc.; 2000.
37. Mitchell E, Noh J, Li S, et al. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022:1754-1768.
38. Jain S, Ma X, Deoras A, Xiang B. Self-consistency for open-ended generations. arXiv, arXiv:230706857, 2023, preprint: not peer reviewed. Published online.
39. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. In: *Proceedings of Conference Advances in Neural Information Processing Systems*. 2023.
40. Song F, Yu B, Li M, et al. Preference ranking optimization for human alignment. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 38. 2024:18990-18998. https://doi.org/10.1609/aaai.v38i17.29865
41. Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. In: *Proceedings of The Eleventh International Conference on Learning Representations*. 2022.
42. Si C, Gan Z, Yang Z, et al. Prompting GPT-3 to be reliable. In: *Proceedings of The Eleventh International Conference on Learning Representations*. 2022.
43. Yukun Z, Lingyong Y, Weiwei S, et al. Improving the robustness of large language models via consistency alignment. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024:8931-8941.
44. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider C, Forte AJ. AI and ethics: a systematic review of the ethical considerations of large language model use in surgery research. *In: Healthcare*. 2024;12(8):Vol:825. MDPI
45. 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):43-49.
46. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Med Internet Res*. 2023;25: e48009.
47. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). arXiv, arXiv:240314473, 2024, preprint: not peer reviewed. Published online.
48. Hanna JJ, Wakene AD, Lehmann CU, Medford RJ. Assessing racial and ethnic bias in text generation for healthcare-related tasks by ChatGPT1. medRxiv. Published online 2023, preprint: not peer reviewed.
49. Rodriguez JA, Alsentzer E, Bates DW. Leveraging large language models to foster equity in healthcare. *J Am Med Inf Assoc*. 2024: ocae055. https://doi.org/10.1093/jamia/ocae055
50. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940.
51. Rane NL, Tawde A, Choudhary SP, Rane J. Contribution and performance of ChatGPT and other large language models (LLM) for scientific and research advancements: a double-edged sword. *Int Res J Mod Eng Technol Sci*. 2023;5(10):875-899.
52. Thapa S, Adhikari S. ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. *Ann Biomed Eng*. 2023;51(12):2647-2651.
53. Luo Y, Zhang J, Fan S, et al. BioMedGPT: open multimodal generative pre-trained transformer for biomedicine. arXiv, arXiv:230809442, 2023, preprint: not peer reviewed. Published online.
54. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. *J Am Med Inf Assoc*. 2024:ocae045. https://doi.org/10.1093/jamia/ocae045
55. Xie Q, Chen Q, Chen A, et al. Me LLaMA: foundation large language models for medical applications. arXiv, arXiv:240212749, 2024, preprint: not peer reviewed. Published online.

Journal of the American Medical Informatics Association, 2024, 31, 1743–1753
https://doi.org/10.1093/jamia/ocae145
Research and Applications