Large Language Models in Neurology Research and Future Practice

Michael F. Romano, MD, PhD, Ludy C. Shih, MD, MSc, Ioannis C. Paschalidis, PhD, Rhoda Au, PhD, and Vijaya B. Kolachalama, PhD

Correspondence
Dr. Kolachalama
vkola@bu.edu

Neurology® 2023;101:1058-1067. doi:10.1212/WNL.0000000000207967

Abstract

Recent advancements in generative artificial intelligence, particularly using large language models (LLMs), are gaining increased public attention. We provide a perspective on the potential of LLMs to analyze enormous amounts of data from medical records and gain insights on specific topics in neurology. In addition, we explore use cases for LLMs, such as early diagnosis, supporting patient and caregivers, and acting as an assistant for clinicians. We point to the potential ethical and technical challenges raised by LLMs, such as concerns about privacy and data security, potential biases in the data for model training, and the need for careful validation of results. Researchers must consider these challenges and take steps to address them to ensure that their work is conducted in a safe and responsible manner. Despite these challenges, LLMs offer promising opportunities for improving care and treatment of various neurologic disorders.

Introduction

Large language models (LLMs) have emerged as a powerful tool for analyzing and interpreting enormous amounts of data. Adding to the fervor is the capacity of LLMs as a form of generative artificial intelligence (AI) able to construct meaningful and contextually appropriate text based on a given prompt, emulating human-like creativity, and reasoning. The excitement and speculation generated by recent reports and media coverage on the potential of LLMs has led to additional questions posed in the public sphere surrounding their appropriate use.[1-3] One of the primary reasons for the surging public interest is the ability of LLMs to generate text that is on-par, if not better than humans, when prompted with questions. For example, GPT-4, OpenAI's latest LLM, has scored high enough to pass all 3 parts of the US Medical Licensing Examination.[4] Such models provide an opportunity to help address existing scientific, clinical, and social needs.

LLMs are deep learning frameworks designed to process natural language text (Table 1 and Table 2 for a glossary of technical terms).[5,6] Unlike more traditional machine learning models, such as naïve Bayes classifiers, which rely on explicit labels ("happy" or "sad"), features (for example, full words or phrases), and rules to identify patterns, LLMs learn to recognize patterns and fill gaps or generate text using deep learning with vast amounts of data. LLMs are typically trained using large text corpora, such as text on the Internet, Wikipedia, books, newspaper articles, and other documents. Once an LLM has been trained, it can be used to perform a variety of tasks, such as language translation, text summarization, and generation of human-like text.

From the Department of Medicine (M.F.R., R.A., V.B.K.), Boston University Chobanian & Avedisian School of Medicine, MA; Department of Radiology and Biomedical Imaging (M.F.R.), University of California, San Francisco; Department of Neurology (L.C.S., R.A.), Boston University Chobanian & Avedisian School of Medicine; Department of Electrical and Computer Engineering (I.C.P.), Division of Systems Engineering, and Department of Biomedical Engineering; Faculty of Computing and Data Sciences (I.C.P., V.B.K.), Boston University; Department of Anatomy and Neurobiology (R.A.); The Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine; Department of Epidemiology, Boston University School of Public Health; Boston University Alzheimer's Disease Research Center (R.A.); and Department of Computer Science (V.B.K.), Boston University, MA.

Go to Neurology.org/N for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

The Article Processing Charge was funded by the authors.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND), which permits downloading and sharing the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

 Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology.

Glossary

AI = artificial intelligence; **IMDRF** = International Medical Device Regulators Forum; **LLM** = Large language model; **NLP** = natural language processing; **PHI** = protected health information; **SaMD** = Software as a medical device.

In the context of research related to neurologic disorders, LLMs could be trained de novo or "fine-tuned" on large clinical data sets to identify patterns and relationships that would be difficult for humans to detect manually.[7-9] This article aims to provide researchers and clinicians with a discussion of this emerging technology and highlighting its potential for study, treatment, and care.

How Large Language Models Work

LLMs are built using neural network architectures that allow them to recognize complex patterns in natural language data and generate realistic text. Although there are several language models published in the literature,[7,10,11] most LLMs use a specific type of neural network called a transformer, which is designed to handle long sequences of text. These neural networks are organized into multiple layers, with each layer consisting of multiple attention heads and feedforward neural networks (see Table 2 for a glossary of these terms). Attention heads are components that decide which parts of the input text the model should focus on when it generates output, helping to understand context and relationships between words. These attention heads consist of matrix multiplications and other mathematical operations. Feedforward neural networks then process the output of the attention heads and compute higher-order relationships between these dependencies and contextual relationships.

To train an LLM (Figure 1), large natural language data sets are used to determine the weights of the neural network. During training, the LLM is presented with input sequences and asked to predict the next word, fill "masked" words, or generate a new sentence based on the input. By minimizing the difference between the predicted output and the actual output, the LLM gradually learns to recognize patterns in the data and generate text that is consistent with the input.

Language models can be trained using different methods,[12-16] such as supervised learning, unsupervised learning, or self-supervised learning. Supervised learning is the most common approach to training language models, where a model is trained on a data set of labeled data and learns to predict the correct label for a given input. When there are no labeled data, then unsupervised learning can be used. An unsupervised learning model could be trained to generate new text by being given a large corpus of text in which it can learn patterns. Self-supervised learning is a newer approach to training language models that has been shown to be effective in learning complex relationships between words and phrases.[17] In this case, the model is trained on a data set that has been artificially labeled, and the model learns to predict a missing label for a given input. For example, a self-supervised learning model could be trained to answer questions by being given a data set of questions and answers. The model would learn to predict the answer to a given question by finding patterns in the data. For LLMs, such as GPT-4, designed to respond well to questions and prompts, the system's output is improved by letting it interact with human testers and applying reinforcement learning techniques.

In addition to the stages presented in Figure 1, LLM development should consider how biases are addressed during the training process. This can be accomplished by, for example, ensuring that the training data are representative of the relevant population. For population-level queries, it is crucial to integrate fairness metrics during the fine-tuning process to assess model performance across different subgroups. Explicit instructions can be given to the model during this phase to avoid bias, thus further promoting unbiased outputs. Improving alignment between model output and the relevant task by iterating on the model's behavior should also be performed. Improving the clarity of guidelines given to human reviewers and developing upgrades to allow users to customize the model's behavior within broad bounds may aid in achieving this alignment. These measures ensure that the model's decisions are interpretable and explainable, allowing us to better understand any underlying bias. Regular audits should also be performed to monitor ongoing model outputs. Overall, these proactive steps need to be integrated to ensure that the LLM development process is mindful of biases and maintains a consistent alignment with human values.

Owing to the ability of these models to process enormous amounts of data, including medical records and patient interviews, and generate high-quality text that accurately reflects the complex symptoms and experiences of patients, LLMs constitute suitable tools for neurologic research and practice. LLM development has evolved over the past decade (Table 3), and the current state-of-the-art models can perform many tasks, including language modeling, text classification, and sentiment analysis.

Language modeling is a fundamental task in natural language processing (NLP) that involves training LLMs to predict the next word in a sentence based on the context of the previous words. This task is often referred to as autoregression and can be performed in 2 ways: left to right or right to left. In left-to-right language modeling, the LLM predicts the next word in a sentence based on the context of the words to its left. By contrast, right-to-left language modeling involves predicting the next word based on the context of the words to its right.

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

Table 1 Types of Large Language Models

Model	Definition
ELMo	ELMo, or Embeddings from Language Models, is a language model that generates contextualized word representations, allowing for improved performance in a range of natural language processing tasks It is a deep contextualized word representation model developed by researchers at the Allen Institute for Artificial Intelligence. It generates word embeddings that capture both syntax and semantics of the input text. Unlike traditional word embeddings, which are fixed and context-independent, ELMo embeddings are dynamic and context-dependent, meaning that they can capture multiple meanings of a word depending on the context in which it is used. ELMo uses a bidirectional long short-term memory (LSTM) network architecture, which allows the model to learn representations of words that consider the context in which they appear. ELMo has been shown to outperform traditional word embeddings on a variety of natural language processing tasks, including sentiment analysis, named entity recognition, and text classification The original reported ELMo model contained 94 million parameters
BERT	BERT, or Bidirectional Encoder Representations from Transformers, is a language model designed for text classification It is a transformer-based language model developed by Google that uses bidirectional training to improve contextual understanding of text. Unlike previous language models, which only used unidirectional training, BERT is trained in both directions of the input text to better capture the context and meaning of the text. BERT is typically trained using a masked language modeling objective and a next sentence prediction objective. It has achieved state-of-the-art results on a wide range of natural language processing tasks, including question answering, sentiment analysis, and language translation The original paper reported 2 models: BERTBASE, which contained 110 million parameters and BERTLARGE contained 340 million parameters. Also, in 2020, NVIDIA released Megatron BERT which contained 3.9 billion parameters, making it the world's largest BERT model at 12x the size of BERTLARGE.
GPT	GPT, or Generative Pretrained Transformer, is a language model designed for natural language processing tasks, such as text generation and question answering It is a large language model developed by OpenAI trained on a massive corpus of text data unsupervised. GPT uses a transformer architecture, which is a type of neural network that is particularly effective at processing sequential data. GPT generates text by taking input sequences and predicting the next word or sequence of words. The model can be fine-tuned on specific tasks, such as text classification or question answering It must be noted that more advanced versions of the GPT model (i.e., GPT-2, GPT-3, and GPT-4) are made available for users The original GPT model (i.e., GPT-1) contained 117 million parameters, the GPT-2 model contained 1.5 billion parameters, and the GPT-3 model, which is an autoregressive language model, contained 175 billion parameters

Each large language model (LLM) listed below has its own unique features and capabilities. These models have revolutionized natural language processing and have enabled researchers and practitioners to develop powerful tools for analyzing and understanding text data.

This function of LLMs can be bidirectional,[5] predicting the next word based on context either to the left or to the right. This flexibility is important when dealing with conversational or narrative data where the meaningful context can come from either direction.

Having learned to predict and fill in "masked" words, LLMs could theoretically aid in communication for those with language impairment due to dementia or a traumatic brain injury. For these persons, this could mean using context to fill in gaps in a patient's narrative caused by memory loss, expressive aphasia, or poor engagement in conversation. This could potentially facilitate better communication between patients and their loved ones or caregivers.

LLMs For Cognitive Assessment and Rehabilitation

LLMs could also be applied to analyze language patterns in patients' spoken or written communication, potentially revealing cognitive shifts or deficits. For instance, by training LLMs on language data collected from patients with Parkinson's disease, at high risk for Huntington Disease, or at high risk for Alzheimer disease, it may become possible to detect subtle variations that evolve gradually, which human observers might overlook. These variations could encompass alterations in word production, vocabulary choice, sentence structure, or the sophistication of concepts over time. Detecting such audio and linguistic changes early on could enable timely intervention and tailored rehabilitation strategies. In a similar vein, LLMs could provide useful tools to augment clinician expertise in identifying language deficits in persons who experienced traumatic brain injury or undergone tumor resection, possibly facilitating more targeted cognitive rehabilitation.

As proof of concept, data sets from 2 recent challenges (ADReSS and ADReSS$_o$) inspired the research community to develop automated methods to analyze speech, acoustic, and linguistic patterns in individuals to detect cognitive changes.[18,19] Valsaraj et al.[20] leveraged pretrained BERT to extract features on the autogenerated transcripts and assess cognitive function. Similar work was performed by Vats et al.,[21] where they used BERT to perform dementia classification. Our own group previously showed that frameworks such as BERT and neural network–based sentence encoding can be used to automatically transcribe digital voice recordings and differentiate cognitively impaired persons from those with normal cognition.[22] Agbavor and Liang[23] similarly leveraged GPT-3 to develop a model to predict dementia in persons using their spontaneous speech.

Cognitive rehabilitation itself could also benefit from language modeling. For example, based on data about a patient's linguistic abilities, an LLM could generate word games or storytelling activities that match the patient's current cognitive level. By tracking the patient's performance over time, the model could adjust the difficulty of the tasks, providing a form of dynamic cognitive stimulation and training. Tasks assessing semantic (category) and phonemic (letter) fluency are commonly used in neuropsychological evaluations for cognitive impairment.[24] In

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

Table 2 Glossary of Technical Terms

Model	Definition
Natural language processing	Often abbreviated as NLP, this is a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language in a valuable and meaningful way
Deep learning	An advanced form of machine learning that involves training neural networks to recognize complex patterns within data through layers of interconnected nodes, where each layer extracts progressively more abstract features
Large language model	An advanced artificial intelligence tool that, having learned from analyzing massive amounts of text data, can generate human-like text based on the context provided
Transformer	This modeling architecture, which was first designed for text data, understands and generates language by comprehending multiple parts of text simultaneously, thereby improving language task performance
Attention	The attention mechanism is a component of a neural network that allows the model to focus on certain parts of the input data more than others, enhancing its ability to understand context and nuances in complex data-like language
Lemmatization and stemming	These techniques are used in natural language processing. Stemming is a method where words are reduced to their base or root form, often leading to grammatically incorrect roots, while lemmatization transforms words to their dictionary form, ensuring linguistic correctness
Autoregression	This is a concept in statistics where current values of a time series are predicted using previous values, serving as a fundamental approach for time-dependent data analysis
Feed-forward network	This is a type of neural network in which information passes from one layer (see: Deep Learning) to a subsequent layer. This contrasts with different types of models, some of which incorporate "loops" where data from subsequent layers is used as input to earlier layers
Reinforcement learning	A type of machine learning where an agent learns to make decisions by taking actions in an environment to maximize a reward signal, progressively improving its behavior based on trial and error

these tasks, individuals are asked to generate as many words as possible from a specific category or starting with a specific letter within a given time. LLMs could be used to automate the analysis of these tasks, providing scores based on not just the number and correctness of words generated but also the uniqueness of the temporal speed variations. This could lead to more objective, reliable, and efficient scoring of these assessments.

The implementation of LLM-based chatbots represents another transformative aspect with exciting potential.[8,25,26] These digital assistants could be programmed to respond to frequently asked questions about their condition, propose various care management options, or even offer emotional support to patients and caregivers.

Electronic Health Record Text Classification

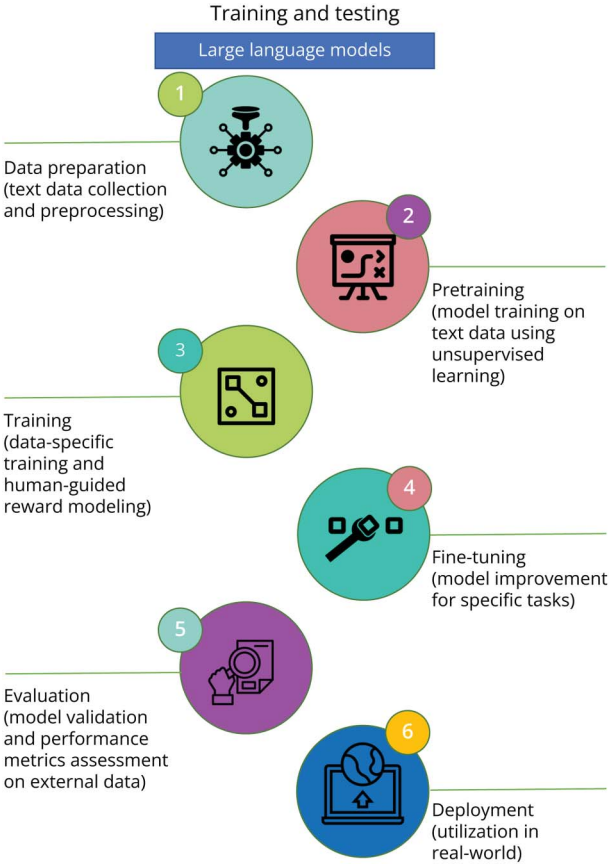
Text classification is a common NLP task where LLMs are trained to classify a given text into categories.[27] This task involves providing a model with a set of text examples and corresponding category labels to learn patterns and relationships between the 2, thereby helping to automate the process of assigning categories to new text data. Named entity recognition is a similar task more restricted to assigning categories to individual "things" within a sentence; for example, the word "Boston" or "London" might be assigned the category "city."

To train an LLM for text classification in a clinical context, text data, such as medical records, patient histories, or clinical trial reports, must be preprocessed. This processing involves tokenizing the text into individual words or subwords, removing irrelevant words, and applying various forms of normalization, such as stemming or lemmatization (Table 2). In brief, stemming and lemmatization are techniques in NLP that reduce words to their root form,[28] with stemming chopping off the ends of words while lemmatization uses vocabulary and morphological analysis to find the base or dictionary form of a word, known as the lemma. Once preprocessed, an LLM is trained to predict a category label, such as a specific neurologic disorder, based on the features present in the text. LLMs can accurately perform text classification because of their capacity to learn intricate representations of text and automatically extract relevant features for classification.

As a potential application of text classification, clinical notes during patient encounters or personal narratives provided by patients about their symptoms could be processed and classified by LLMs to identify patterns of text that might correlate with specific neurologic conditions. This could help automate the categorization of new patient information into relevant classes such as 'migraine' and 'Parkinson,' enabling more efficient analysis of patient data and serving as a clinical assistant. For example, Gehrmann et al.[29] analyzed discharge summaries using NLP and convolutional neural networks, finding that they were able to categorize respective persons as having "chronic pain," "advanced cancer," or "advanced lung disease," among others. This could be extended to relevant neurologic diagnoses or categories. In the context of neuroimaging, LLMs can identify noteworthy information in radiology reports in emergency department settings.[30] In cases of stroke, for instance, where timely intervention is critical and patient communication may be impaired due to aphasia or other neurologic deficits, an LLM could serve as an effective tool to flag essential neuroimaging findings for providers.

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

Figure 1 Schematic of the General Process for Training and Testing a Large Language Model



(1) Data preparation: This step involves collection and preprocessing a large corpora of text data to be used for model training. (2) Pretraining: This step involves training the model on the large corpus of text data using unsupervised learning frameworks. The goal is to predict missing words or predict the next word in a sentence, given the previous words. (3) Training: Here, the model is further trained using a more specific data set, often involving human supervision. This stage includes the process of "reward modeling" where the model generates a set of potential responses, and human reviewers rank them. The model uses this ranking to generate responses in the future. This stage is crucial to ensuring the alignment of the model's outputs with human values and instructions. (4) Fine-tuning: This step involves improving the trained model on a smaller, labeled data set for a specific task. (5) Testing and evaluation: This key step involves validation of the fine-tuned model on an external (i.e., separate) data set and evaluate the model's performance on a specific task. Metrics, such as accuracy, precision, recall, and F1-score, can be used for model evaluation. (6) Deployment: If the validated model meets the desired performance criteria, then it can be used to perform the specific task in a real-world setting.

When applied to large-scale data sets, such as electronic health records (EHRs) or databases of scientific literature, LLMs could improve classification accuracy and help streamline the process of clinical observational research. For example, Fernandes et al.[31] demonstrated that an NLP algorithm was able to assign neurologic disability outcomes after intensive care unit hospitalization, based on free-text clinical notes. In another study, Xie and coauthors used 3 different LLMs (BERT, RoBERTa, and Bio_ClinicalBERT) to comb through clinical notes and determine whether and how frequently patients had seizures.[32] Furthermore, LLMs' abilities to learn from multiple languages allows for cross-lingual text classification,[33] which could enable the classification of neurologic data regardless of the language, thereby benefitting global neurologic research and patient care.

Text classification could also assist interpretation of neuropsychological tests, brain imaging, neuropathology, and neurophysiology studies, such as electroencephalography and electromyography/nerve conduction study reports. It could help to automatically categorize parts of these reports into clinically relevant predefined classes (e.g., normal/abnormal findings, presence/absence of certain key terms). In addition, it could potentially identify trends across multiple assessments, such as during a complicated or prolonged hospitalization, providing a clearer picture of a patient's clinical trajectory over time. Overall, using LLMs for text classification could enhance the speed and efficiency of processing patient data.

Sentiment Analysis

Sentiment analysis involves training an LLM to identify an underlying sentiment or emotion expressed in text.[34] Given the increase in patient provider communication occurring outside the direct face-to-face encounter, such as through patient portal messaging, language processing methods such as sentiment analysis could offer insights into patients' subjective experiences and emotional states, which could be profoundly affected by neurologic impairment and the understanding of which is critical to managing these conditions. The goal of sentiment analysis is to automatically identify the polarity of a text, such as a patient message, voice recording, or video recording, which could be, for example, positive, negative, or neutral. Such analysis could provide crucial insights into patients' psychological well-being, their experiences with various treatments, or the impact of their neurologic condition on their day-to-day life and signal the need for prioritizing dedicated behavioral and mental health resources for the patient.

To train an LLM for sentiment analysis, a labeled data set is required, consisting of text data and its corresponding sentiment labels. Creating an effective data set demands a careful, domain-specific approach. The labeling process in neurology, for example, should ideally involve annotators skilled not only in language but also familiar with the intricacies of neurologic disorders. They would assign sentiment labels to textual data, reflecting a range of emotional responses that are common to patients experiencing neurologic conditions and to their caregivers. Creating a suitable data set for neurology-centric sentiment analysis also calls for balance and representation. As with other common LLM tasks, labelling should cover a variety of neurologic conditions, treatments, and patient-caregiver interactions to avoid model bias and accurately capture the breadth of sentiment in this field. While some generic resources, such as the PhysioBank,[35] can provide a base, researchers should also look for neurology-specific data. The LLM would then be trained to identify patterns and relationships between text samples and their respective sentiment labels.

When trained on large-scale data sets such as databases of patient narratives or clinical communication, LLMs can learn complex

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

Table 3 Timeline of the Development of LLMs

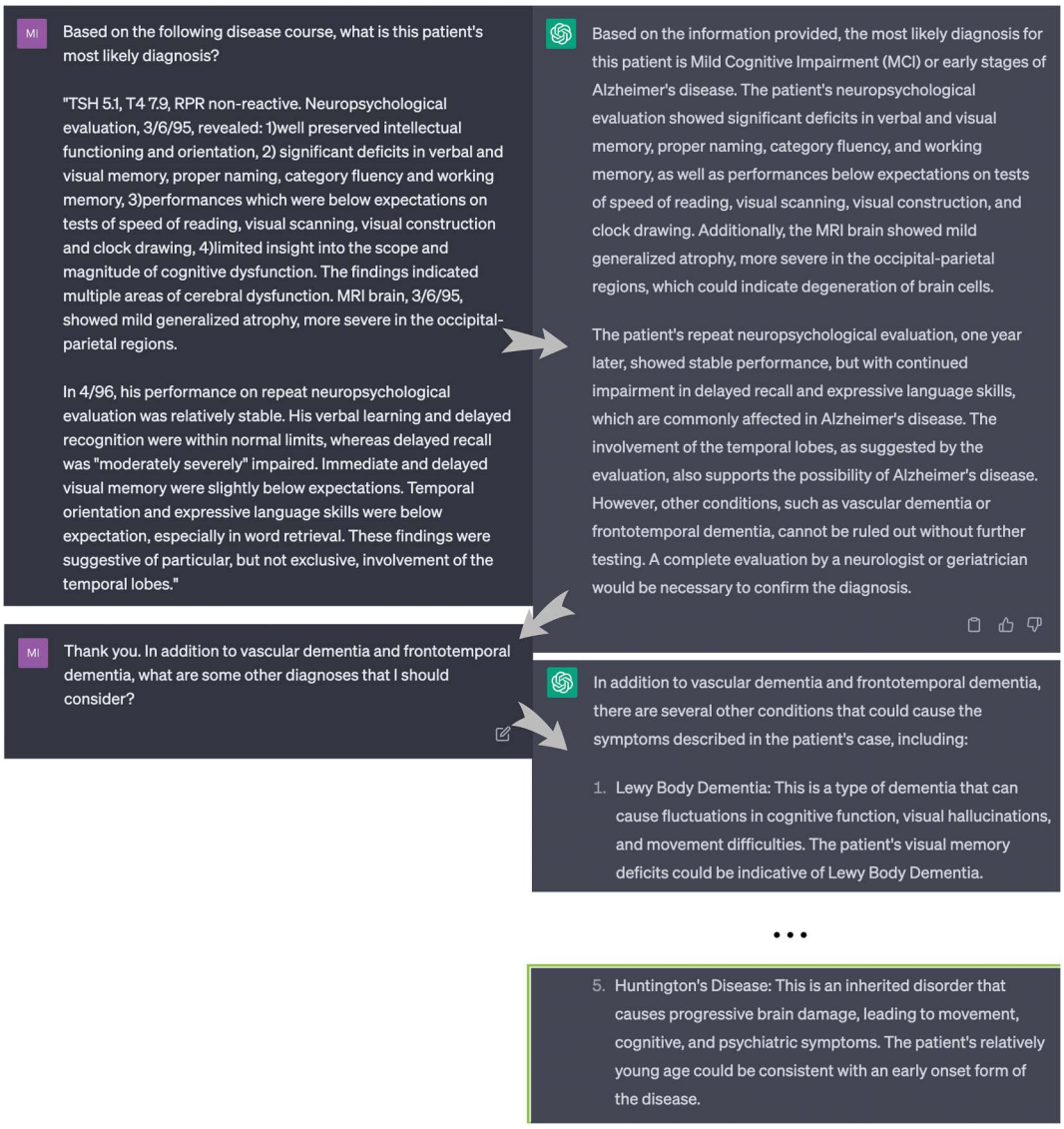
2011: A Recurrent Neural Network–based Language Model (RNNLM) was proposed, which served as an important predecessor to many modern LLMs
2013: Google Brain researchers trained a neural network to learn word representations from enormous amounts of text data, known as word2vec. This represented a significant advance in natural language processing and helped to pave the way for LLMs
2013: Stanford NLP Group released the Stanford Parser, an open-source software that provides grammatical analysis tools to researchers
2014: Sequence to Sequence Learning with Neural Networks was published, which laid the foundation for many of the developments in LLMs
2015: OpenAI was founded to develop advanced AI models safely and responsibly. This included the development of LLMs, such as GPT-2 and GPT-3
2016: Google's Parsey McParseface, an open-source syntactic parser, was a significant contribution to the field
2017: Transformer was introduced, which is a model architecture that provided the groundwork for many subsequent LLMs due to its efficient handling of long-range dependencies
2018: ELMo (Embeddings from Language Models) was introduced by researchers at Allen Institute for Artificial Intelligence. ELMo uses a deep, bidirectional LSTM to generate word embeddings that can capture both syntax and semantics
2018: The Bidirectional Encoder Representations from Transformers (BERT) model was introduced by Google researchers, representing a significant advance in the development of LLMs. BERT is capable of bidirectional training, allowing it to better understand the context and meaning of language
2018: ULMFiT (Universal Language Model Fine-tuning) was introduced, marking an important milestone in the efficient use of transfer learning in NLP.
2018: OpenAI released the first version of GPT, or Generative Pretrained Transformer, which used unsupervised learning to generate text
2019: Facebook AI Research introduced RoBERTa, or Robustly Optimized BERT approach, which was designed to improve the accuracy and robustness of BERT by modifying the training process
2019: OpenAI introduced GPT-2, a highly advanced LLM capable of generating realistic and coherent human-like text. Owing to concerns about potential technology misuse, OpenAI first released a limited version of the model
2019: Facebook AI Research introduced XLM (Cross-Lingual Language Model), a pretrained LLM capable of understanding multiple languages
2019: Hugging Face, an AI community-driven company, released the transformers library, an open-source resource providing pretrained models and tools to the NLP research community
2019: Text-to-text transformer was released, which allowed reframing NLP tasks into a unified text-to-text format where the input and output is always text
2020: ELECTRA was released, which is a pretraining approach which trains a transformer model to distinguish "real" input tokens vs "fake" input tokens generated by another transformer model
2020: Google's Meena and Facebook's Blender chatbots were introduced, both of which are large-scale conversational AI models
2020: Developed by NVIDIA, Megatron is designed to scale up models such as GPT and BERT for model size, data set size, and the amount of compute
2020: OpenAI released GPT-3, the most advanced LLM to date, with 175 billion parameters. GPT-3 has demonstrated impressive performance on a range of natural language processing tasks, such as language translation and question answering
2022: The BigScience consortium released BLOOM, which an open-science LLM trained on 46 natural languages and 13 programming languages
2023: OpenAI released GPT-4, endorsing a performance on the Uniform Bar Examination in the 90th percentile and a performance on the Medical Knowledge Self-Assessment Program in the 75th percentile
2023: Falcon LLM was released by the Technology Innovation Institute, which is an open-source model for commercial and research use
2023: Pythia, which is a suite of 16 models, was released to enable scientific research on openly accessible and transparently trained LLMs
2023: Med-PaLM and Med-PaLM2 were released by Google, designed to provide helpful long-form answers health-related questions
2023: Meta released LLaMA and LLaMA2, an open-source LLM attempting to democratize access worldwide

A summarization of the key aspects that underscore the progress made on the development of LLMs over the past few years. While we attempted to cover most of the recent advancements, this list is by no means fully exhaustive.

text representations, capturing subtle nuances and contextual information essential to understanding patients' sentiments. Moreover, LLMs can identify the sentiment of individual words or phrases within a text, providing deeper insights into the emotions and opinions expressed by patients or their caregivers. For instance, this could enable the identification of specific symptoms or side effects that cause substantial distress or therapies that lead to positive emotional responses. Consider a scenario where LLMs are trained on a diverse range of text data from individuals with depression. These models could learn to detect subtle changes in language patterns, sentiment, and expression that correlate with the progression or severity of depression, provide insights into early signs of worsening depression, or even predict episodes of heightened distress that signal to care teams that close follow-up is needed.

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

Figure 2 Conversation With GPT



Demonstration of ChatGPT's ability to summarize a clinical note and generate plausible differential diagnoses. Highlighted in green is an example of a "hallucination," a current limitation of the technology, where in this case, ChatGPT "hallucinates" that the patient has a "young age," although the age of the patient in the clinical vignette was never specified. The clinical summary is abbreviated from a publicly available resource (mtsamples.com).

Additional Work on the Use of Language Models in Neurology

The use of large language models in clinical neurology is a developing field. Most work to date related to language modeling has focused on report generation, clinical documentation, and patient health record analysis using more established approaches, such as text mining and NLP, with additional examples detailed above.[36-38] Although it would be beyond the scope of this article to delve into all related research, we highly recommend that readers peruse several review papers for a more comprehensive understanding of this rapidly evolving but young field.[38-44] In fact, in a recently published abstract in *Neurology*, Lefkovitz et al.[44] highlighted gaps in neurologic NLP research, with few to no studies in certain neurologic disorders. We offer additional areas of neurology in which LLMs may be helpful in eAppendix 1 (links.lww.com/WNL/D193).

Technical and Ethical Challenges

Need to Ameliorate Bias

The potential for bias in LLMs presents unique technical and ethical challenges. The complex and heterogeneous nature of neurologic disorders necessitates that any model used in this field be trained on diverse, representative, and consistently collected data to keep from propagating health care disparities between different groups of people. Most notably, sampling bias has the potential to propagate preexisting health care disparities given that neurologic conditions manifest

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

differently across a range of demographics, including age, sex, and ethnic groups. For instance, Alzheimer disease presents with varying symptoms and progression rates that differ between individuals and demographic groups. For classification, if an LLM was trained predominantly on data from older adults older than 65 years, its utility in diagnosing early-onset Alzheimer disease could be compromised. Strategies, such as oversampling, underrepresented patient groups, and using debiasing techniques during model training could be used to counter this bias. Training transparency and interpretability techniques could help clinicians understand the reasoning behind LLM recommendations, enhancing trust and clinical adoption.

In the realm of sentiment analysis, sentiment misinterpretation could have significant clinical repercussions such as leading to inappropriate interventions or treatments. To mitigate this, rigorous training of LLMs on diverse language patterns, including nuances in emotional expression across different patient groups, is vital. In addition, integrating feedback loops with health care professionals to validate and fine-tune sentiment predictions could enhance accuracy.

Measurement bias can occur due to the variety of tools and methods used in neurologic assessments, such as cognitive tests, neuroimaging techniques, and neurologic examinations. Data collected from these disparate sources might introduce inconsistency and variability. This bias can be minimized by using standardized protocols for data collection and incorporating a wide range of data sources to train the model. Confirmation and reporting biases pose significant risks in the context of neurology because of the subjectivity involved in assessing symptoms, such as pain, fatigue, or cognitive changes. Overrepresentation or underrepresentation of these symptoms in the training data could result in a skewed model that fails to accurately predict these aspects in patients. Given these biases' potential to affect an LLM's output and thus potentially affect patient care, researchers must generate rigorous clinical evidence through controlled studies assessing the accuracy, benefits, risks, and adverse events of incorporating LLMs in neurology. Furthermore, neurologists must be aware of an LLM's limitations and understand its generalizability across different neurologic conditions and patient demographics. It is crucial for them to approach LLMs as an aid rather than a replacement for their clinical judgment and expertise.

Need for Careful Technical Validation

The inherent complexity of LLMs can pose challenges in neurology. For instance, addressing "hallucinations,"[45] where a model might generate significant errors, is critical in neurology where precision in data interpretation is paramount. An example of a "hallucination" from ChatGPT is shown in Figure 2, where it inaccurately assigns a "young age" to a patient based on a clinical note fragment without any age given.

To address this challenge, LLMs for neurology ought to undergo careful technical validation to ensure that they are safe and helpful for their intended uses. This validation should include not only generalizable methods such as cross-validation and independent testing on data sets with varied demographics but also tests relevant to specific neurologic disorders. For instance, a model's ability to accurately predict dementia onset from clinical notes or neuroimaging data should be tested using data not involved in the model's training. Furthermore, as detailed above, careful attention should be given to potential biases or limitations in the training data. If the training data overrepresent certain demographics, the model's output may not be accurate or reliable when applied to underrepresented groups. Rigorous methods should be used to mitigate bias during data collection and curation, and the model's performance should be tested across diverse demographics.

Need to Preserve Privacy and Maintain Data Security

Machine learning model development in general, and specifically LLMs, presents significant privacy and data security concerns that must be addressed to protect the rights and confidentiality of study participants and patients.[46] In addition to privacy concerns, there are data security concerns associated with the use of LLMs in clinical practice. LLMs are complex models that require significant computational resources to train and run. Researchers must ensure that appropriate data security measures are in place to protect the models and the data used to train them, such as secure cloud-based storage and access controls, and to prevent data breaches, such as regular security audits, data encryption, and secure data transmission. To address these concerns, researchers must ensure that they comply with relevant data protection laws and regulations, such as the General Data Protection Regulation in the European Union and the Health Insurance Portability and Accountability Act in the United States.

Several aspects regarding the usage of LLMs in practice must be carefully considered to ensure that the research is conducted in a responsible and transparent manner, particularly with respect to the principle of autonomy, and the right to decide how one's protected health information (PHI) is used by LLMs. The screening of large EHR databases may require special notification to patients who are vulnerable and may require a waiver of consent granted by institutional review boards to use LLMs to screen EHR data. Inherent to this task is ensuring the privacy and confidentiality of the data being used to train the models. PHI must be carefully protected to avoid any unintended harm or discrimination, particularly against individuals who may have impairment or disability.

Another challenge associated with the use of LLMs in neurology is obtaining proper informed consent from patients or their legally authorized representatives, including in situations when the initial consent to the use of PHI data is given when the individual is cognitively intact, and only later becoming cognitively impaired. It is critical that the evaluation of institutional review boards be included when making determinations about appropriateness of consent, particularly in the context of the evolution of consent as new scientific advances continue to emerge.

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

Role of Regulation

Federal regulations could serve as a useful adjunct to technical and clinician expertise in addressing the limitations and challenges of LLMs. Many areas of regulation lie outside the scope of this article, although there are several regulatory issues that are particularly important with respect to neurology and more broadly, medicine, that pertain to the technical and ethical challenges we raise.[19,47,48]

Software as a medical device (SaMD) is defined by the International Medical Device Regulators Forum (IMDRF), as software that is not embedded within hardware and which performs medical tasks. Therefore, many medical LLMs would fall under this umbrella. As Gilbert et al.[19] note, even LLM chatbots used for clinical decision support could be considered medical devices. Under the IMDRF framework, adopted as guidance by the FDA, SaMD would need to meet 3 standards during clinical evaluation. These include (1) that there must be an association between SaMD output and the relevant clinical condition; (2) that an input generates "accurate, reliable, and precise" output; and (3) that the output achieves the desired goal in the population of interest. Any regulatory efforts should keep these standards in mind.

Bazoukis et al.[49] introduce the idea of incorporating algorithm auditing to augmented intelligence models. Adapted to LLMs specifically, this could involve labelling models with segments of the population on which a particular model may be less effective or even untested. In addition, regulatory bodies could introduce mandated testing of LLMs on a private validation data set with demographics representative of the general population or specific marginalized populations. These steps, mandating testing on standardized data sets and labelling algorithms with expected performance on different population segments, could help to address bias and technological validation during an initial approval process and would help LLMs meet the standards set by the IMDRF framework for clinical evaluation.

Regulation may also be helpful in safeguarding patient data. There need to be specific timelines for removal of patient data from models and data sets and rules regarding the use of generative models pretrained on a patient's data if a respective patient wants their information removed. A balance between feasibility and patient safety must be navigated carefully, and new techniques may need to be developed to hasten this process.

Conclusion

LLMs offer opportunities within the realm of neurology, promising to bolster diagnostic accuracy, expedite early interventions, and unravel new biomarkers and therapeutic pathways. LLMs can be valuable educational assets for patients and caregivers, elucidating complex neurologic conditions and treatments in a digestible manner. The utility of LLMs could also extend to the enhancement of diagnostic precision, as demonstrated by the potential identification of subtle linguistic changes indicative of early-stage cognitive impairment, through patient narrative processing. Nonetheless, it is crucial to address a few challenges, such as standardizing the training of LLMs to minimize biases, safeguarding patient privacy, and ensuring technical validation. As we navigate these challenges, there is a need for interdisciplinary collaboration, encompassing computer scientists, neuroscientists, ethicists, and clinicians. We call for further research efforts in areas, such as neurology-specific data annotation, bias mitigation in LLM application to neurologic conditions, and the development of more transparent models capable of delivering clinically meaningful insights. Despite these limitations, LLMs can serve as powerful agents of change within neurology. We must harness our collective knowledge and resources to foster research collaboration, develop unbiased and transparent LLMs, and undertake initiatives to bridge existing knowledge gaps. It is through these concerted efforts that we will move closer to fully unlocking the potential of LLMs in improving accurate diagnosis and treatment of neurologic disorders.

Study Funding

This project was supported by grants from the Karen Toffler Charitable Trust, the American Heart Association (20SFRN35460031), the NIH (RF1-AG062109, R01-HL159620, R21-CA253498, R43-DK134273, RF1-AG072654, U19-AG068753, and P30-AG013846), the National Science Foundation under grants CCF-2200052, DMS-1664644, and IIS-1914792, and a pilot award from the National Institute on Aging's Artificial Intelligence and Technology Collaboratories (AITC) for Aging Research program.

Disclosure

The authors report no relevant disclosures. Go to Neurology.org/N for full disclosures.

Publication History

Received by *Neurology* June 1, 2023. Accepted in final form September 6, 2023. Submitted and externally peer reviewed. The handling editor was Editor-in-Chief José Merino, MD, MPhil, FAAN.

Appendix Authors

Name	Location	Contribution
Michael F. Romano, MD, PhD	Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, MA; Department of Radiology and Biomedical Imaging, University of California, San Francisco	Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; analysis or interpretation of data
Ludy C. Shih, MD, MSc	Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, MA	Drafting/revision of the manuscript for content, including medical writing for content; analysis or interpretation of data

Downloaded from https://www.neurology.org by Boston University on 26 January 2024

Appendix *(continued)*

Name	Location	Contribution
Ioannis C. Paschalidis, PhD	Department of Electrical and Computer Engineering, Division of Systems Engineering, and Department of Biomedical Engineering; Faculty of Computing and Data Sciences, Boston University, MA	Drafting/revision of the manuscript for content, including medical writing for content; analysis or interpretation of data
Rhoda Au, PhD	Department of Medicine; Department of Neurology; Department of Anatomy and Neurobiology; The Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine; Department of Epidemiology, Boston University School of Public Health; Boston University Alzheimer's Disease Research Center, MA	Drafting/revision of the manuscript for content, including medical writing for content; analysis or interpretation of data
Vijaya B. Kolachalama, PhD	Department of Medicine, Boston University Chobanian & Avedisian School of Medicine; Faculty of Computing and Data Sciences; Department of Computer Science, Boston University, MA	Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data

References

1. The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit Health*. 2023;5(3):e102. doi:10.1016/S2589-7500(23)00023-7.
2. Sanderson K. GPT-4 is here: what scientists think. *Nature*. 2023;615(7954):773. doi:10.1038/d41586-023-00816-5
3. Arora A, Arora A. The promise of large language models in health care. *Lancet*. 2023;401(10377):641. doi:10.1016/S0140-6736(23)00216-7
4. Nori H, King N, McKinney SM, Carignan D, Horvitz E. *Capabilities of GPT-4 on Medical Challenge Problems*; 2023.
5. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*. 2018.
6. OpenAI. GPT-4 Techincal report. *arXiv*. 2023.
7. Singhal K, Azizi S, Tu T, et al. Publisher correction: large language models encode clinical knowledge. *Nature*. 2023;620(7973):E19. doi:10.1038/s41586-023-06455-0
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8
9. Hao B, Zhu H, Paschalidis I. Enhancing clinical BERT embedding using a biomedical knowledge base. Proceedings of the 28th International Conference on Computational Linguistics 2020:657-661.
10. Zhao WX, Zhou K, Li J, et al. A survey of large language models. *arXiv*. 2023.
11. Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models. *J Biomed Inform*. 2022;126:103982. doi:10.1016/j.jbi.2021.103982
12. Lai TM, Zhai C, Ji H. KEBLM: knowledge-enhanced biomedical language models. *J Biomed Inform*. 2023;143:104392. doi:10.1016/j.jbi.2023.104392
13. Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artif Intelligence Med*. 2021;117:102083. doi:10.1016/j.artmed.2021.102083
14. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform*. 2020;8(3):e17984. doi:10.2196/17984
15. Laparra E, Mascio A, Velupillai S, Miller T. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearb Med Inform*. 2021;30(1):239-244. doi:10.1055/s-0041-1726522
16. Savova GK, Danciu I, Alamudun F, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer Res*. 2019;79(21):5463-5470. doi:10.1158/0008-5472.CAN-19-0579
17. Chen Y-P, Lo Y-H, Lai F, Huang C-H. Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: algorithm development and validation study. *J Med Internet Res*. 2021;23(1):e25113. doi:10.2196/25113
18. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Detecting cognitive decline using speech only: the ADReSSo Challenge. *medRxiv*. 2021.
19. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. *Nat Med*. 2023;29:2396-2398. doi:10.1038/s41591-023-02412-6
20. Valsaraj A, Madala I, Garg N, Baths V. Alzheimer's dementia detection using acoustic & linguistic features and pre-trained BERT. 2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI). 2021: 171-175.
21. Vats NA, Yadavalli A, Gurugubelli K, Vuppala AK. Acoustic features, bert model and their complementary nature for Alzheimer's dementia detection. 2021 Thirteenth International Conference on Contemporary Computing (IC3-2021). 2021: 267-272.
22. Amini S, Hao B, Zhang L, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. *Alzheimers Dement*. 2022. doi:10.1002/alz.12721
23. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. *PLOS Digit Health*. 2022;1(12):e0000168. doi:10.1371/journal.pdig.0000168
24. Nutterupham K, Saykin A, Rabin L, et al. Verbal fluency performance in amnestic MCI and older adults with cognitive complaints. *Arch Clin Neuropsychol*. 2008;23(3):229-241. doi:10.1016/j.acn.2008.01.005
25. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838
26. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184
27. Dogra V, Verma S, Kavita, et al. A complete process of text classification system usingstate-of-the-artNLP models. *Comput Intell Neurosci*. 2022;2022:1883698. doi:10.1155/2022/1883698
28. Chai KEK, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. *J Am Med Inform Assoc*. 2013;20(5):980-985. doi:10.1136/amiajnl-2012-001409
29. Gehrmann S, Dernoncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One*. 2018;13(2):e0192360. doi:10.1371/journal.pone.0192360
30. Evans CS, Dorris HD, Kane MT, et al. A natural language processing and machine learning approach to identification of incidental radiology findings in trauma patients discharged from the emergency department. *Ann Emerg Med*. 2023;81(3):262-269. doi:10.1016/j.annemergmed.2022.08.450
31. Fernandes MB, Valizadeh N, Alabsi HS, et al. Classification of neurologic outcomes from medical notes using natural language processing. *Expert Syst Appl*. 2023;214:119171. doi:10.1016/j.eswa.2022.119171
32. Xie K, Gallagher RS, Conrad EC, et al. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. *J Am Med Inform Assoc*. 2022;29(5):873-881. doi:10.1093/jamia/ocac018
33. Almagro M, Martínez R, Montalvo S, Fresno V. A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation. *J Biomed Inform*. 2019;94:103207. doi:10.1016/j.jbi.2019.103207
34. Denecke K, Reichenpfader D. Sentiment analysis of clinical narratives: a scoping review. *J Biomed Inform*. 2023;140:104336. doi:10.1016/j.jbi.2023.104336
35. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*. 2000;101(23):101. doi:10.1161/01.cir.101.23.e215
36. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243. doi:10.1136/svn-2017-000101
37. Aneja S, Chang E, Omuro A. Applications of artificial intelligence in neuro-oncology. *Curr Opin Neurol*. 2019;32(6):850-856. doi:10.1097/WCO.0000000000000761
38. Yew ANJ, Schraagen M, Otte WM, van Diessen E. Transforming epilepsy research: a systematic review on natural language processing applications. *Epilepsia*. 2023;64(2):292-305. doi:10.1111/epi.17474
39. Crema C, Attardi G, Sartiano D, Redolfi A. Natural language processing in clinical neuroscience and psychiatry: a review. *Front Psychiatry*. 2022;13:946387. doi:10.3389/fpsyt.2022.946387
40. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. *Int J Methods Psychiatr Res*. 2016;25(2):86-100. doi:10.1002/mpr.1481
41. 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
42. Rezaii N, Wolff P, Price BH. Natural language processing in psychiatry: the promises and perils of a transformative approach. *Br J Psychiatry*. 2022:1-3. doi:10.1192/bjp.2021.188
43. Baldassano SN, Hill CE, Shankar A, Bernabei J, Khankhanian P, Litt B. Big data in status epilepticus. *Epilepsy Behav*. 2019;101(Pt B):106457. doi:10.1016/j.yebeh.2019.106457
44. Lefkovitz I, Walsh S, Blank L, Jette N, Kummer B. *Direct Clinical Applications of Natural Language Processing in Common Neurological Disorders: A Systematic Review (PS-4.005)*; 2023.
45. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? *Crit Care* 2023;27(1):148. doi:10.1186/s13054-023-04425-6
46. Thapa C, Camtepe S. Precision health data: requirements, challenges and existing techniques for data security and privacy. *Comput Biol Med*. 2021;129:104130. doi:10.1016/j.compbiomed.2020.104130
47. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. doi:10.1038/s41746-023-00873-0
48. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA*. 2023;330(4):315-316. doi:10.1001/jama.2023.9651
49. Bazoukis G, Hall J, Loscalzo J, Antman EM, Fuster V, Armoundas AA. The inclusion of augmented intelligence in medicine: a framework for successful implementation. *Cell Rep Med*. 2022;3(1):100485. doi:10.1016/j.xcrm.2021.100485

Downloaded from https://www.neurology.org by Boston University on 26 January 2024