Received: 13 October 2023 | Revised: 30 December 2023 | Accepted: 19 January 2024

DOI: 10.1111/epi.17907

Epilepsia®

CRITICAL REVIEW

Potential merits and flaws of large language models in epilepsy care: A critical review

Eric van Diessen[1,2] | **Ramon A. van Amerongen**[3] | **Maeike Zijlmans**[4,5] | **Willem M. Otte**[1]

[1]Department of Child Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands

[2]Department of Pediatrics, Franciscus Gasthuis & Vlietland, Rotterdam, The Netherlands

[3]Faculty of Science, Bioinformatics and Biocomplexity, Utrecht University, Utrecht, The Netherlands

[4]Department of Neurology and Neurosurgery, UMC Utrecht Brain Center, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands

[5]Stichting Epilepsie Instellingen Nederland, Heemstede, The Netherlands

Correspondence
Eric van Diessen, Department of Child Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht and Utrecht University, Room KG 01.310.0, P.O. Box 85090, Utrecht 3508 AB, The Netherlands.
Email: e.vandiessen-3@umcutrecht.nl

Funding information
EWUU grant "AI for Health" (2021)

Abstract

The current pace of development and applications of large language models (LLMs) is unprecedented and will impact future medical care significantly. In this critical review, we provide the background to better understand these novel artificial intelligence (AI) models and how LLMs can be of future use in the daily care of people with epilepsy. Considering the importance of clinical history taking in diagnosing and monitoring epilepsy—combined with the established use of electronic health records—a great potential exists to integrate LLMs in epilepsy care. We present the current available LLM studies in epilepsy. Furthermore, we highlight and compare the most commonly used LLMs and elaborate on how these models can be applied in epilepsy. We further discuss important drawbacks and risks of LLMs, and we provide recommendations for overcoming these limitations.

KEYWORDS

diagnosing epilepsy, large language models, natural language processing

1 | INTRODUCTION

Natural language processing (NLP) is a form of artificial intelligence (AI) that focuses on the computational analysis of spoken and written language to retrieve relevant information or patterns. NLP is used increasingly in the medical domain to automatically extract data from electronic health records (EHRs), scientific articles, patients blog posts, audio recordings, or medical guidelines. AI gained momentum by introducing a new and generic model architecture in the seminal paper "Attention is all you need," allowing pre-trained large language models (LLMs).[1] After the implementation of ChatGPT by OpenAI as the first, large ready-to-use

LLM for the general public in November 2022,[2] an ever-increasing number of users were seen, and applications were developed, including within the medical domain.[3-5] LLMs can be best appreciated as a new class of NLP, typically trained on large data sets and used for both "passive" tasks (e.g., information extraction and pattern detection in texts) and more "active" tasks (e.g., text generation and transformation).[6,7]

In this review, we critically appraise the development of LLMs and introduce this rapidly expanding field for clinician–scientists who are considering using these models in clinical or research settings. This hands-on overview aims to improve understanding and integration of LLMs. In addition to an introduction to the development history, key ingredients, and limitations, we present current LLM studies in epilepsy research and provide examples for future LLM applications in epilepsy care. Given the importance of "language input" for diagnosis, treatment evaluation, and patient management, LLMs seem particularly appealing for improving epilepsy care. Apart from a recently published commentary-including clarifying examples of how LLM can be used in epilepsy care—little of the current LLM literature has focused on the possible merits and flaws in epilepsy research.[8] Finally, we reflect on how LLMs differ from traditional NLP applications, as we recently reviewed.[9]

2 | DEVELOPMENT OF LARGE LANGUAGE MODELS

2.1 | Historical overview: From text mining to transformers

2.1.1 | What milestones have led to the development of LLMs?

NLP is rooted in 1950, with the development of the infamous "Turing test." In this thought experiment, the British computer scientist Alan Turing assessed whether a machine could engage in a natural conversation indistinguishable from human conversation. Turing's machine is generally conceived as a first attempt to generate natural language semi-automatically.[10] It was not until the early 1990s that NLP algorithms further evolved with the availability of more powerful computers and access to large textual data sets. Machine learning (ML) models drove these NLP algorithms and could deduce linguistic patterns from text without following rules to break down text.[11] Typically, these algorithms were trained on existing textual data to learn relevant patterns, and consequently used to make predictions on new, unseen data. Initially, these machine-based NLP algorithms were trained in a 'supervised' manner: models trained on texts that were manually labeled with linguistic properties to produce similar linguistic labels for unseen text. Later, deep learning and artificial neural networks were introduced to improve model performance and generalizability to large texts.[12] This allowed the representation of words and sentences in a vector space, enabling the expression of the proximity of linguistic properties (e.g., semantics and syntax) in this space.[13] The subsequent introduction of "self-supervised" LLMs was a game-changer for the field: it allowed models to learn patterns from large quantities of non-labeled text through self-adjustment and obviated human supervision.[1] Because LLMs are generalizable to a variety of language comprehension and generation tasks, interest in these models increased rapidly, leading to the development of more sophisticated models primarily by large tech companies, such as Google, OpenAI, and Meta.

Key points

- Although large language models (LLMs) are applied increasingly in medical care, few studies have applied this strategy in epilepsy care.
- Epilepsy care could benefit from integrating LLMs to accelerate diagnosis and to facilitate clinical evaluation and patient counseling.
- A stepwise introduction of LLMs into epilepsy care is proposed to avoid misinterpretation and inappropriate use in daily practice.

2.2 | Mechanisms

2.2.1 | What are the key ingredients of LLMs?

Most current LLMs have a so-called "transformer architecture." Tokens are the elementary LLM's input and output language units. Tokens are generally—but not exclusively—restricted to words, subphrases, or punctuation marks. This tokenization of language is needed for LLMs to operate efficiently and capture combined words, intentionality, and grammar. For the readability of this section, we refer to "words" when we technically mean tokens. To read a text, a clinician extracts information from the direct context of words: preceding and following words, sentences, and even the entire paragraph. Transformers do something similar, but fast and on a massive scale. The mechanism allows the model to focus on some words more than others, depending on what makes sense in the context. This allows the model to properly process individual words and the relationships between words, and the overall meaning of a sentence or paragraph. More technically, given a text as input, LLMs will first internally represent words into a vector by which a linguistic relation (e.g., semantic, syntax, grammar) between word positions is expressed as a numerical weight. This representation of a word position computed as the weighted relation to all other words is a vector.[11] What makes attention so crucial for transformer-based LLMs? Attention enables the model to focus on different parts of a sentence and simultaneously skip both the non-informative information and extraction of relevant information from each computational step (or "layer") in the model, rather than looking at the last computation step only. By doing so, meaningful information typically stored in deeper model layers is used for improved model output.[13] A self-attention mechanism was added subsequently to reduce the computational price and make LLMs less prone to forgetting information that is typically stored in the model (Figure 1; Box 1). The Bidirectional Encoder Representations from Transformers (BERT) model, developed by Google, was one of the first LLMs to incorporate this transformer architecture and attention mechanisms.[14] Later models adopted this architecture, such as the Generative Pre-Trained Transformers (GPT) developed by OpenAI.[15]

Compared to BERT, the self-attention mechanisms of GPT are slightly modified: each word position is related only to all preceding positions but not to succeeding ones. Because succeeding positions are ignored (masked), self-attention in GPTs is called "masked self-attention." GPTs are, therefore, "autoregressive": they use only past positions to predict the coming "masked" word, as opposed to BERT, which uses past and succeeding positions because words are masked at random positions in a sentence. Several LLMs have been published, and although output generation has considerably improved, some limitations remain. Most notably, an autoregressive LLM like GPT still has difficulties completing complex (non-sequential) tasks. A linear way of thinking prevents these models from accessing information recursively, that is, making, planning, and memorizing necessary intermediate steps. However, the most recent versions of GPT incorporate new strategies for recursively accessing information.[16]

2.2.2 | What is the scope of current LLM applications?

The current interest in LLM has resulted in the development of many different, (non) open-source models

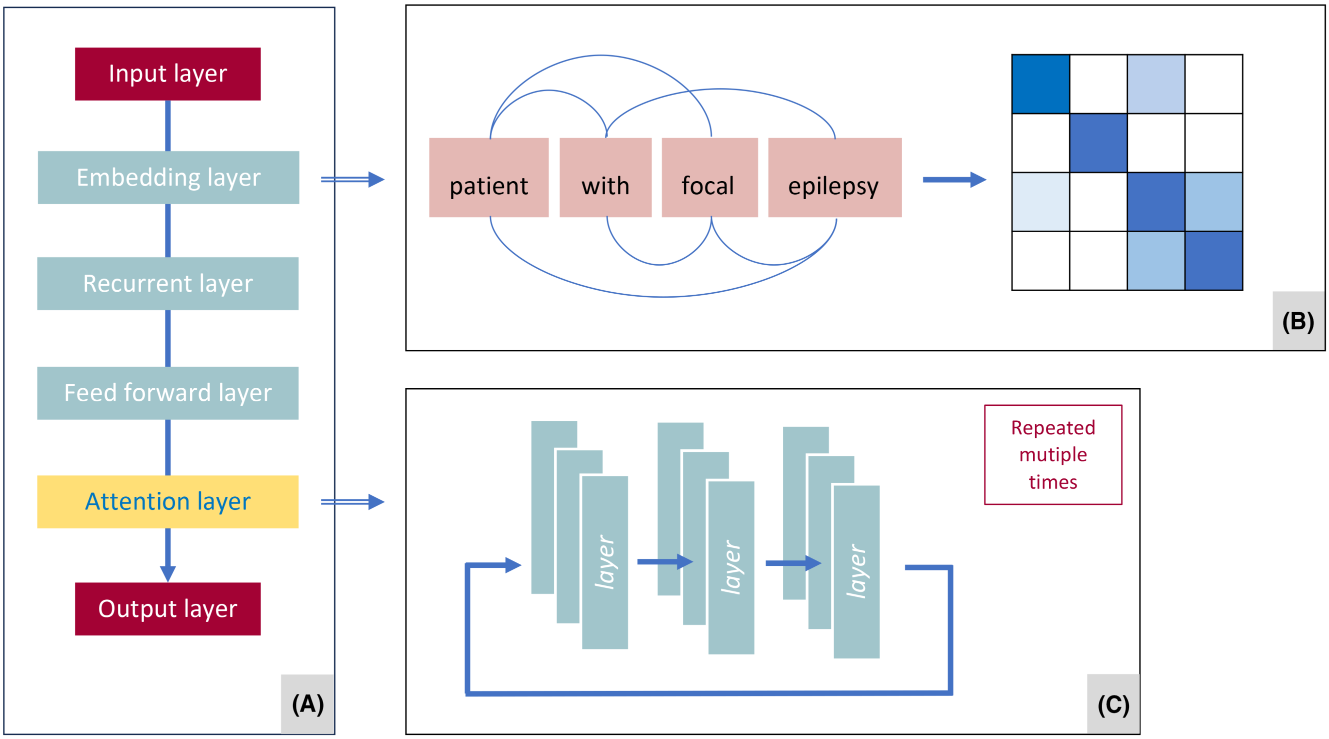

FIGURE 1 Simplified representation of the most essential LLM components. (A) Schematic illustration of a large language model (LLM), including the most commonly used model layers. (B) In the embedding layer, the linguistic relation between words (or tokens)—curved lines—are represented in a vector in which a numerical weight is given to every relation (the darker the color, the more related the words are). (C) An essential component of LLM is attention (including self-attention and/or masked attention, depending on the type of model used). Attention enables retrieval of relevant information for the requested output from any layer in the model.

BOX 1 A glossary of basic LLM concepts

Transformer: dominant architecture of current LLM in which different attention mechanisms are crucial for model output.

Attention: a key component of transformer-based LLMs that allows extraction of information typically stored in different model layers. Self-attention is a mechanism in which attention is focused selectively on the input sequence used for the model's output.

Pretraining: refers to the initial phase of model training. During this phase, the model learns to predict the next word in a sentence or fill in missing words based on the context of the surrounding words and capture the nuances of language and understand various linguistic patterns. Large-scale pretraining enables the model to grasp general language features and information. After the pretraining phase, the model can be fine-tuned on a smaller domain-specific data set to target a specialized task, such as medical text analysis.

Layer: an LLM is typically composed of different layers. Each layer is a computational step with unique properties to process input text and generate output.

Token: the basic units that LLM uses to process language. Tokenization is splitting texts into smaller units that an LLM can process. Tokens generally correspond to words, but can also be subphrases, word syllables, or punctuation marks depending on the type and size of the model. The memory of an LLM is often expressed in token limit, referring to the maximum number of tokens that can be processed.

Vector: a mathematical way to represent words (or tokens) with numerical weights that express the relation with other words in the given text.

Prompting: instructions or questions a user provides to shape a model's output relevant to the requested task. Prompt engineering is the process of constructing effective prompts for LLM. Prompt strategy depends on the desired output and the model's available (labeled) data.

over the last few years. In most cases, these models are pretrained on large quantities of text and typically differ in size and complexity, as expressed in their number of parameters (Table 1). More parameters correspond to superior learning and task-generation performances but come with a higher computational price. Recent studies suggest that increasing data heterogeneity and training length could improve LLM performance with smaller data sets. For example, the Large Language Model Meta AI (LLaMA), developed by Meta, outperforms GPT-3 on several benchmarks despite fewer parameters (13 billion vs 175 billion).[17] This improved performance of smaller, often open-sourced LLMs is particularly appealing for resource-poor and underfunded research areas. A model can be improved to achieve the desired output in several ways.[15] For domain-specific LLMs, one can either fine-tune generic pretrained models on domain-specific texts[18,19] or exclusively build an LLM on non-annotated domain-specific data-like medical texts.[20] The latter strategy is applied increasingly as it avoids the laborious work of supervised data labeling. In task-agnostic LLMs—models that are generic and not pretrained on domain-specific input—human feedback can be used for model alignment and reduction of toxic output.[15]

2.2.3 | What is the influence of prompting on model output?

In addition to model characteristics, the output quality of LLMs depends on instructions or questions provided by their users. Minor differences in prompts, or the use of different versions of LLMs, may result in different outputs and could potentially limit the reproducibility of these models in (clinical) research.[21] Prompt engineering is essential for gaining a desirable model output. For example, a zero-shot approach (i.e., a single question) can be used when a straightforward answer or task is required, and no specific or up-to-date information is needed. In cases of more complex tasks, it might be necessary to guide the model by combining some examples (shots) of input–output for the desired output. This so-called, few-shot prompting improves model performance by using the capacity of LLMs to learn repetitive patterns effectively, especially in situations with limited trained labeled data.[22]

2.2.4 | What kind of medical tasks can LLMs perform?

From the capacity of LLMs to predict a next word in line, a wide range of text-related tasks emerge: text generation, transformation, and summarization but also information retrieval, conversation, and inferring (or combination). A clear and understandable prompt—a written (or spoken) instruction—is pivotal and crucial for output quality.[23] More recently we have seen LLMs performing increasingly complex tasks in the medical domain. For example,

TABLE 1 Overview of current LLMs models.

Model	Base model	Parameters #	Domain	Open source	Release date
GPT-2	–	1.5B	General	Yes	11/2019
Bio_ClinicalBERT	BERT	.34B	Medical	Yes	2020
GPT-3	–	175B	General	No	6/2020
GPT-3.5 (ChatGPT)	–	1.3B – 175B	General	No	11/2022
Claude	–	52B	General	No	12/2021
BioMedLM	–	2.7B	Medical	Yes	12/2022
LLaMA	–	7.0B – 65B	General	Yes	2/2023
GPT-4	–	–	General	No	3/2023
Cerebras-GPT	GPT-3	.1B – 13B	General	Yes	3/2023
ChatDoctor	LLaMA	7.0B	Medical	Yes	3/2023
Dolly 2.0	–	12B	General	Yes	4/2023
Open-Assistant	–	13B	General	Yes	4/2023
PaLM 2	–	340B	General	No	5/2023
Med-PaLM 2	PaLM 2	340B	Medical	No	5/2023

Note: When a base model is given, the domain fine-tuned model will have the same number of parameters. Different versions of a model are indicated by multiple parameter numbers. "B" stands for "billion." Current models are often (derivatives of) Bidirectional Encoder Representations from Transformers (BERT), Pathways Language Model (PaLM), Large Language Model Meta AI (LLaMA), and Generative Pretrained Transformers (GPT).

an instruction-tuned Pathways Language Model (PaLM) variant could answer (complex) medical questions well but could not outperform clinicians on all topics.[24] To better understand and answer medical challenges, current work has been dedicated to improving models by incorporating not just language data, but also multimodal input from images, sensors, wearables, genomics, and so on, into a single model.[25] With this, newer versions of Med-PaLM will be able to synthesize and communicate information from ancillary investigations like magnetic resonance imaging (MRI) scans, x-rays, electrocorticography (ECG), electroencephalography (EEG), mammograms, and more, to assist clinicians in diagnosing patients with a multimodal input. The multimodal input fits the latest tendency to fuse research fields, which were traditionally separate in approach and technique. Computer vision, speech recognition, robotics, image generation, speech synthesis, and even music generation use pretrained language models with their data converted into one form of text or another.[26]

3 | APPLICATIONS OF LLMs IN EPILEPSY CARE

3.1 | Clinical challenges in epilepsy care

Recent progress in epilepsy classification and improved ancillary investigations have significantly narrowed the diagnostic gap in epilepsy. Efforts by the International League Against Epilepsy (ILAE) to refine the classification of epilepsies and seizures are invaluable in evaluating individuals who are suspected of having seizures.[27] The updated ILAE classification provides a clinical framework to better understand, counsel, and manage people with epilepsy (PWE). An intrinsic limitation of the current framework is the clinical variants and atypical presentations that do not adhere to a specific seizure category or epilepsy classification.[28] Similarly, the ongoing development of neuroinflammatory, genetic, and structural biomarkers is pushing the diagnostic and treatment opportunities forward,[29] but neglects an important source of data: the language used in the clinical setting by PWE. Language is widely acknowledged as an indispensable source of information in diagnosing epilepsy—clinicians take history and distill relevant clinical variables from a patient's narrative.[30,31] The upsurge of NLP to systematically process textual data provides a unique opportunity to use this information source for clinical purposes, especially in more complex clinical cases.[32-34] We recently reviewed the current literature in epilepsy research on NLP applications and summarized the evidence for improving epilepsy diagnosis and management, in addition to the previously mentioned efforts.[9] Considering the ever-growing use of "eHealth" in epilepsy care, including EHRs, online patient communities, and the possibilities for digital interaction with clinicians, the availability of textual data will only

BOX 2 Current LLM studies in epilepsy

The number of research articles on LLM applications in epilepsy is limited but is expected to increase rapidly. A PubMed literature search performed on December 6, 2023, revealed five original peer-reviewed research studies; two were published in the last 2 months (Appendix S1). Methodological details, used models, and outcome measures are presented (Table 2). Because two studies from the same research group greatly overlapped in methodology, data, hypothesis, and outcome measure,[19] we presented only the most recent and clinically relevant study.[40]

increase.[35] Yet, language yields rather "noisy" data in which relevant symptoms are sometimes difficult to apprehend. LLMs open new opportunities to use raw textual data for improving epilepsy care and research but is used only limitedly in epilepsy research (Box 2). The next sections integrate a few examples and elaborate on possible research opportunities.

3.2 | Bridging the current challenges with LLMs

3.2.1 | Diagnosing epilepsy

A recent review discusses the automatic extraction of relevant clinical information from epilepsy-related EHRs, including diagnostic characteristics.[36] The potential promise of automatic extraction anticipates the power and ease of LLMs to do so. Even traditional ML approaches have shown potential. For example, Connolly and colleagues classified an NLP model to classify generalized, focal, or unclassified epilepsy in beyond-hospital patient notes.[37] Fonferko-Shadrach and colleagues developed an extraction of epilepsy-related clinical text software system to identify information—demographics, diagnosis, pathology type, and medication use—from clinic-free texts, including letters.[38] More recently a feasible NLP approach, with higher validated accuracies, was used to identify seizure types and frequencies from EHRs.[39] Two recently published diagnostic studies in epilepsy have implemented LLMs systematically in large data sets.[40,41] Both studies applied an LLM approach to extract clinical information on epilepsy outcomes from unstructured clinical notes. Xie and colleagues showed that an adapted BERT model outperformed more traditional (rule-based)

TABLE 2 Overview of current LLM studies in epilepsy.

Author	Topic	Model used	Type and size of data set	Main outcome	Performance
Beaulieu-Jones et al.[41]	Management of first-time seizure-like events in children	Clinical-longformer model	14021 EHRs (control cohort 15062 EHRs)	Predicting seizure recurrence based on physician's notes	F-score: .83–.90
Kim et al.[53]	Providing accurate educational information on epilepsy	GPT-3.5 and GPT-4	57 medical questions	Correct response to an epilepsy-related question	Correct responses: 70%–98%
Xie et al.[40]	Extracting epilepsy outcome from EHRs	Bio_ClinicalBERT and RoBERTa	55630 clinic notes	Accurate classification of seizure freedom	F-score: .88
Wu et al.[52]	Prediction of seizures in EEG	BERT	7 long-term intracranial EEG recordings	Epileptic seizure prediction in time-frequency domain	Sensitivity and FPR: .86; .18/h

Note: Bio_ClinicalBERT and RoBERTa are models derived from BERT but with additional clinical pre-training and more data, respectively.

Abbreviation: FPR, false-positive rate.

NLP methods in extracting information from clinical notes on seizure outcomes.[40] Beaulieu-Jones and colleagues used a clinical-longformer model, similar to a BERT model, but with improved performances due to a more-efficient, self-attention mechanism.[42] In doing so the authors revealed that unstructured clinical notes contain potentially valuable data that are typically neglected in conventional case evaluation for seizure recurrence.[41,43] This ability of LLMs to extract information from clinical consultation, typically overlooked in conventional diagnostic decision models, has not yet been fully explored. For example, experienced and expressed emotions from clinicians and patients influence clinical decision-making,[44] but are largely neglected as contributing variables in the diagnostic process. Patient conversations typically contain these data, which can be extracted with audio recordings and written transcripts. Pevy and colleagues used this approach and revealed that formulation efforts (i.e., hesitations, reformulations and syntactic repairs)—retrieved with an open-source NLP toolkit from spoken patient conversations—could help epileptologists to differentiate between epileptic and non-epileptic seizures.[45] Considering the unique features and superior performance of LLMs in retrieving hidden information from text, LLMs can identify overlooked or misunderstood clinical information to improve diagnostic performances of current clinical decision models.

Finally, with the availability of (future) multi-model driven LLMs,[25] the ability and performance to diagnose PWE by integrating all information available during consultations will increase. This availability could include language sources like referral letters, patient questionnaires, audio-recorded anamnesis and hetero-anamnesis, EHR content, and (home) videos of seizures. Diagnostics such as genomics, imaging, raw EEG time series, and laboratory tests can be added to the equation.

3.2.2 | Personalized care and stratifying patient groups at risk

An ongoing challenge in epilepsy management is to tailor treatment after clarifying the diagnosis. Information on treatment response, side-effects, and drug-resistance for the individual patient—combined with the doctor's knowledge on efficacy and potential side effects—would ideally guide these decisions.[46] Even more, individual patient characteristics such as medical history, personality traits, genomics, and epilepsy traits may help to select the optimal therapy.[47,48] A recent study reviewed the current available clinical decision models in epilepsy management and concluded that the utility of these models remained to be determined.[49] Modest sample size, lack of external validation, and statistical challenges contribute to the eventual clinical integration. From a data perspective, an additional limitation of these decision models is the restricted use of structured texts or pre-defined clinical variables. Advanced NLP models like LLMs, however, allow the exploration of unstructured text to retrieve relevant clinical information. For example, Vulpius and colleagues showed that validating epilepsy diagnosis and stratifying patients into different epilepsy types from (unstructured) EHRs was feasible, using a named-entity recognition (NER) algorithm. By applying this text-mining approach, the algorithm was able to identify a cohort of PWE with a false discovery rate dropping to 4% in International Classification of Disease, Tenth Revision (ICD-10)–registered epilepsy patients and assign focal or generalized epilepsy type for 92% of those patients with an unspecified type.[50] A forthcoming step in this process is forecasting comorbidity, treatment complications, and disease outcomes based on available data. Domain-specific LLMs such as Foresight, a GPT-based model fine-tuned with information from EHRs, are now becoming available for medical forecasting in the individual patient.[51] Models like Foresight are particularly appealing for improving follow-up of complex cases of PWE, as they can extract longitudinal clinical information from unstructured EHRs. Temporal modeling of available (unstructured) EHRs could help predict therapy response and forecast refractory epilepsies and epilepsy-related comorbidities such as cognitive or behavioral deficits at later stages. This temporal profiling may lead to preventive actions and improved tailored counseling. It is notable that the application of LLM is not restricted to language. Wu and colleagues showed that with some model adjustments, EEG time-series can be used in a BERT model to predict seizures in time-frequency domain.[52] This explorative study may set an example in the epilepsy research field for integrating LLM models with neurophysiological measurements aiming to forecast seizure risks in patients and/or outcomes of surgical and therapeutic interventions.

3.2.3 | Improving patient knowledge and compliance

Non-adherence, intentional or non-intentional, is a severe problem in epilepsy care associated with increased mortality, morbidity, and health care costs.[54,55] Education, behavioral, and mixed interventions are ways to improve adherence in PWEs.[56] Time limitations in the outpatient department and/or non-availability of an epilepsy nurse consultant are restrictions to implement face-to-face consultation. To this end, virtual interactive sessions can help improve compliance and the avoidance of seizure triggers.[57] Epilepsy-specific, LLM-trained chatbots would ideally fit in this regimen, being always available for PWE. Questions could range from lifestyle oriented (e.g., when am I allowed to drive again? What sports are considered as high risk?) therapy related (e.g., what do I have to do when I missed one dose of anti-seizure medication? Are my current complaints (known) side-effects?) to prognosis (e.g., when will I be declared cured?). Despite the appealing deployment of LLM-trained chatbots, the quality and applicability will depend on the quality of the underlying model and the information provided. This is a crucial issue, and further consideration is a necessity prior to implementation. Furthermore, LLMs could help to personalize current patient brochures to better highlight relevant complications based on the available clinical data, such as EEG outcomes (photosensitivity), type of genetic mutation (genotype–phenotype), choice of therapy (antiseizure medication, ketogenic diet, or vagus nerve stimulation), or comorbidities (organ-specific or development). In a recent study, Kim and colleagues evaluated the quality output of ChatGPT in answering commonly asked questions by PWE. The authors concluded that ChatGPT is a reliable tool for providing epilepsy-related information, especially on lifestyle-related issues, and could assist the epileptologist in the counseling and education of PWE.[52,53]

3.2.4 | Outsourcing time-consuming clinical work

In addition to involvement in diagnosis, treatment, and patient education, LLMs can also take over more routine tasks of the clinician. Like other physicians, clinical epileptologists or neurologists seeing PWE are coping with an increasing administrative load.[58] Because LLMs excel in generating and transforming texts, they can assist in drafting medical letters. Creating textual (or graphic) patient information materials is a more complex task to improve personalized communication between involved (health) parties.[4] Given the current limitations of LLMs, final output should be supervised and controlled by the responsible clinician.

4 | DISCUSSION

We highlighted and introduced the most commonly used LLMs and elaborated on the challenges and opportunities of integrating generative AI models in epilepsy care and research. This critical review does not discuss the methodological issues of LLM studies systematically, which could have resulted in selection and information bias.

4.1 | Current limitations of LLMs

4.1.1 | What are the common pitfalls and biases of LLMs in the medical domain?

Like any other AI model, LLMs are not infallible, and errors can arise for various reasons (Table 3). A particular challenge that current models face is to critically apprehend information, especially when tasks (or prompts) are more complex. This challenge could result in the interweaving of errors and biases, and the use of faulty information, leading to incorrect conclusions by its users, especially when overly relying on these models (i.e., automation bias). Dratsch and colleagues demonstrated that this automation bias would likely influence radiologists' performance when using automated decision-making software, especially when inexperienced.[59] It is not unlikely that a similar situation can arise when integrating LLMs into epilepsy care. Boßelmann and colleagues provided two illustrative clinical examples of epilepsy, in which ChatGPT provided incorrect suggestions for epilepsy surgery and related a genetic variance incorrectly to epileptogenesis.[8] Erroneous suggestions of LLMs can thus lead to inappropriate treatment strategies and incorrect diagnosis, particularly when the (professional) user is inexperienced or unable to critically apprehend the premises on which the model output is made. A solution is adding a probability score to the model output to clarify for the user the level of accuracy, ideally accompanied by relevant references. This also holds for the deployment of LLM-trained chatbots.

The output of current LLMs often appears overconfident due to two different phenomena: models are prone to "hallucinate" and "hedge." Most LLMs are trained on a fixed data set at a certain time point with no access to recent and up-to-date information. Inaccessibility to recent information or overrepresentation of outdated information can result in the fabrication of incorrect but plausible-sounding answers (i.e., hallucinations).[60,61] Hedging is the tendency of LLMs to give long-winded answers when trying to nuance the outcome and could lead to automation bias because users are more likely to trust weighted answers.[3,18] Overconfidence of LLMs can be alleviated partially by improving the calibration of future models in which a better agreement is reached between the predicted and observed correctness of the model output, but this is currently limited.[62] More recent ChatGPT versions already exhibit a reduced tendency

TABLE 3 List of LLM limitations and solutions.

Limitation	Description	Solution(s)
Limited world knowledge	LLMs do not know of recent events or expert topics	Inserting relevant knowledge in prompt Access to external search engines Fine-tuning to new data
Knowledge bias	Preference by LLMs for (incorrect) information because of, e.g., volume in training data	See "Limited world knowledge" Better balanced training data
Hallucinations	Fabrication of nonsensical and incorrect information by LLMs	See "Limited world knowledge" Fine-tuning to reduce tendency for hallucinations
Hedging	LLMs giving long-winded answers for nuanced topics	See "Limited world knowledge" Fine-tuning to reduce tendency for hedging
Limited context window	LLMs forget previous input/output over longer conversations	
Input sensitivity	Small differences in input can lead to different outputs by LLMs	Prompt-engineering Feedback through prompts
Instruction refusal	LLMs may refuse to follow instructions in prompt	Prompt engineering Feedback through prompts Fine-tuning to better follow instructions
Difficulty with complex tasks	LLMs may have difficulty with mathematical or other complex tasks	Integration with other tools Feedback through prompts
Small working memory	LLMs cannot retain much information while "thinking"	Output intermediate steps
Autoregressive property	GPT models cannot reevaluate previous output because of the forward nature of word prediction	
Social bias	LLMs may perpetuate harmful biases related to race, gender, disability, and so on, from their training data	Filtering of biased content from training data Fine-tuning
Accessibility bias (language)	Limited LLM access for certain demographic groups because of lower performance in certain languages	Including more content of other languages in training data
Accessibility bias (commercialization)	Limited LLM access for certain demographic groups because of commercialization	Open-source initiatives that provide data and LLMs
Toxicity	LLMs can produce toxic content that is prompted deliberately or non-deliberately by the user	Filtering of biased content from training data Fine-tuning External systems for censoring output
Censoring	Censoring output could remove harmless output and ignore harmful output	

Note: The limitations and solutions listed here are non-exhaustive and may not apply to all LLM models. In general, many of these limitations will likely be addressed by future generations of LLMs.

to hallucinate and hedge.[2] From the user's perspective, several actions can be taken: give access to relevant information through an external search engine, fine-tune models to domain-specific, validated data sets, or provide relevant information with a prompt.[16] It is important to note that when consecutive prompts are given, LLMs tend to forget information from previous inputs. As a result, applicability in longer medical conversations may be limited. These challenges are particularly crucial to consider in epilepsy care. PWE typically use different sources of information to inform themselves, but misinformation on epilepsy is abundant, especially on the non-(medically) moderated internet, including social media.[63] Because generic LLMs, such as ChatGPT, typically use all available information for output generation, misinformation can be easily integrated into its responses. An approach to overcome this limitation in future epilepsy-specific LLM development is to build an epilepsy-data library of medically reliable sources.[20] Accuracy assessment is pivotal for LLM studies, albeit sometimes sparsely described. Because generative LLMs essentially learn probabilistic associations between words, it is important to understand how different factors contribute to inaccuracy. A lack of recent content or topic-related data affects pre-training and model output and calls for model fine-tuning and external validation.[24] In a study on predicting seizure recurrence in children, Beaulieu-Jones and colleagues, therefore, used a clinical-oriented LLM, supplied with extracted data from clinical notes, and tested model performance in an external cohort.[41]

4.1.2 | Is censoring harmful output effective and needed?

LLMs are prone to perpetuate biases that are presented in training data. In addition to knowledge bias (Table 3), LLMs are susceptible to social biases. Underrepresentation of minorities and specific groups of people in training data could reinforce social biases related to gender, race, religion, and disability.[64,65] An example of representational harm of specific groups is the tendency of (older) LLMs to unequally present gender representation among different occupations.[66] Recently studies have been published on different mitigation strategies to improve inclusivity and diversity in LLM.[67] Allocation harm—referring to the unfair distribution of resources and opportunities—can potentially arise when LLMs are used in decision-making that directly affects people, such as in health care or career opportunities.[62,65] Venkit and colleagues demonstrated that LLMs like BERT and GPT-2 negatively scored sentences if they contained disability-related words in the context of employment, being homebound, self-care, and physical abilities.[64] PWE—who face disability-related stigma in various contexts—could be particularly at risk for negative influence from this allocation harm. LLMs exhibit lower performance when prompted in certain languages because some languages are overrepresented in the training data. This performance discrepancy may lead to accessibility bias between people of different nationalities. Similarly, the performance of LLMs might be affected by formulation differences in a prompt due to age, educational background, or cognitive abilities.[65] In the case of epilepsy, with a disproportional disease burden in people with cognitive impairments or in resource-poor areas, this could strengthen the knowledge gap and access to a potential source of valuable information.[68] One solution is to use the quality of the prompt and/or retrieved information as a proxy for the users' cognitive abilities and level of knowledge on epilepsy and adjust the output accordingly to improve readability and comprehension.

The issues raised contribute to a so-called alignment problem: do LLMs (continue) to align with human interests and values? As LLMs continue to learn and adapt unpredictably once deployed, anticipating their behavior might become challenging.[69] This also includes the emergent abilities of larger LLMs, which cannot merely be extrapolated from performance of smaller models.[70] Both the alignment problems and "emergent risks" are currently addressed in safety discussions on LLMs and call for at least some form of control or censoring of output. Several options exist, including some form of human intervention, for example, built-in ethical constraints, validating and verifying output generation, or switch-off procedures.[71] Ethical and legal considerations need further elaboration to ensure a solid integration into medical care.[72] Counterfactual fairness has been proposed as a built-in ethical criterion to evaluate the fairness of models' output. For example, what would happen if the information was requested by an individual with a different background or (medical) situation?[73] Also of concern medical data need protection from unauthorized access. Recent studies have showcased unintentional and unauthorized output of LLMs due to memorized information including personal data, after a so-called prompt injection attack.[74] Without thorough investigation and installment of preventive measures, pretraining on epilepsy text files with personal data is dangerous because models are vulnerable to leaks of sensitive information. Compliance with existing regulations is essential (e.g., Accountability Act, Medical Device and General Data Protection Regulation).[75] Guidelines are needed for clinical personnel before using LLM-driven medical tools in clinical decision-making.

4.2 | Future developments and alternatives for LLMs

Large tech companies continue to release improved versions of LLMs that typically deal with the limitations we have discussed. Next to increased model parameters and output accuracy that we previously discussed (Table 1), token limit in LLMs continues to be improved. First-generation LLMs typically contained a limit of 512 tokens, thereby restricting prompt input and model output. Recent versions of LLaMA and ChatGPT extend token size up to 32 k, which allows modeling of a complete patient track record. Another hurdle to overcome is generalizability to languages other than English. Models using cross-lingual, multitask fine-tuning, excellently perform tasks in languages to which they were never exposed and outperformed previous models that relied on language-specific transfer learning.[76,77] In addition, we expect a wide range of commercial activities from smaller commercial players to assist users in the best possible way. Assisted prompting, stepwise output generation, and coding are illustrative applications of recently developed LLMs.[78] Despite benchmark analyses in comparisons with older or competing LLMs, we expect that most of these models will be (at least partially) non-open sourced. Apart from preventing open science, this may limit potential censoring and adaptation for the public. We await more initiatives of the larger tech players in the field to launch open-source models such as LLaMA and Open-Assistant, as this would also accelerate our understanding of how LLMs work.[17,79]

A potential application in the medical domain of LLMs is integration into clinical decision-support systems.[8] LLMs could combine their medical content and capabilities to extract and analyze patient data to provide clinical summaries, draft patient letters, and provide (supportive) advice on differential diagnosis and treatment options. A large health care software company is piloting the integration of LLM in hospitals across the United States.[72] This integration may contribute to the future challenge of more (complex) health demands, fewer people employed in health care, and improved health care in resource-poor settings. In the case of epilepsy, this model could be of added value by evaluating different treatment options, epilepsy surgery planning, or complex genetic testing in search of an etiological (genetic) origin. Given the complexity of most clinical questions—and the current challenges in output generation of LLMs—it is not likely that these models will soon function as completely autonomous agents. Finally, problems due to alignment and the emergent abilities inherent to LLMs call for exploring complementary and alternative approaches. A recent example is cognitive emulation,* which aims to build a more understandable and controllable model that follows a more human way of reasoning and handling failure. Current efforts include a model architecture based on the "thinking fast and slow theory" to constrain AI systems to make decisions in a constrained, but more human-like and understandable environment.[80]

5 | CONCLUSION

We provide an overview of LLMs' current opportunities, challenges, and pitfalls. Although we used the field of epilepsy as an example, these insights can be easily adapted to any other field in clinical (neuro)sciences. Similarly, points raised in other domains advancing the integration of LLMs in clinical care—such as radiology—should be considered a valuable source of information.[4] A particular challenge here is the pace of model development. Current developers are improving model performance with an unprecedented speed, providing potential users with a variety of new options on an almost monthly basis. Most of these LLMs are closed-sourced applications, so a thorough (academic) understanding remains limited. Recent jailbreak attempts have provided a better understanding of the robustness, reliability, and working mechanism of some LLMs,[81-83] but in-depth information on vector construction and attention attribution remains partly unclear. Therefore, future debates in the scientific community should include whether these models are falsifiable.

The era of NLP has provided valuable tools for deriving information from unstructured text that can assist the clinician in daily practice.[9,33,34] With the generative possibilities of modern LLMs, new possibilities arise to further improve epilepsy care. Careful use and knowledge of possible biases are continuous challenges, especially for PWE, with limited possibilities for weighing the model's output. And yet, when used appropriately, LLMs provide an excellent opportunity for providing better patient education and self-management behavior for PWE, especially in areas with limited access to health care. A proposition for the stepwise introduction of LLMs is to start with low-stake situations (i.e., summarizing patient information, generating standard clinical content) rather than more complex clinical cases that include therapeutic advice.[8] Finally, throughout this review, we mentioned some considerations for careful use of current LLMs for clinical assistance. We expect more-robust, domain-specific specialization of open-sourced medical LLMs soon, enabling further integration of LLMs in epilepsy care and research after careful ethical consideration and legal approval.[84]

AUTHOR CONTRIBUTIONS

Ev.D. and W.M.O. initiated the review. Ev.D., Rv.A., M.Z., and W.M.O. contributed to drafting and finalizing the manuscript.

ACKNOWLEDGMENTS

This research was supported by the EWUU grant "AI for Health" (2021). We thank the anonymous reviewers for constructive feedback.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest. We confirm that we have read the journal's position on issues involved in ethical publication and that this report is consistent with those guidelines.

*https://www.conjecture.dev/coem2/.

ORCID

Eric van Diessen https://orcid.org/0000-0002-7773-1990
Maeike Zijlmans https://orcid.org/0000-0003-1258-5678
Willem M. Otte https://orcid.org/0000-0003-1511-6834

REFERENCES

1. Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention is all you need. Adv Neural Inf Proces Syst. 2017;30.
2. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Proces Syst. 2022;35:27730–44.
3. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96.
4. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large Language models are double-edged swords. Radiology. 2023;307(2):e230163.
5. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health. 2023;5(4):e179–e181.
6. Zhou M, Duan N, Liu S, Shum HY. Progress in neural NLP: modeling, learning, and reasoning. Engineering. 2020;6(3):275–90.
7. Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of medicine. J Biomed Inform. 2013;46(5):765–73.
8. Boßelmann CM, Leu C, Lal D. Are AI language models such as ChatGPT ready to improve the care of individuals with epilepsy? Epilepsia. 2023;64(5):1195–9.
9. 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.
10. Turing AM. Computing machinery and intelligence. Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer. Netherlands: Springer; 2009. p. 23–65.
11. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: a methodology review. J Biomed Inform. 2020;109:103526.
12. Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, et al. A survey on deep learning. ACM Computing Surveys (CSUR). 2018;51(5):1–36.
13. Manning CD. Human Language Understanding & Reasoning. Daedalus. 2022;151(2):127–38.
14. Devlin J, Chang M-W, Lee K, Google KT, Language AI. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018 ArXiv preprint arXiv:181004805.
15. Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018.
16. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023 ArXiv preprint arXiv:230312712.
17. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. LLaMA: Open and Efficient Foundation Language Models. 2023 ArXiv preprint arXiv:230213971.
18. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards Expert-Level Medical Question Answering with Large Language Models. 2023 ArXiv preprint arXiv:230509617.
19. Xie K, Gallagher RS, Conrad EC, Garrick CO, Baldassano SN, Bernabei JM, 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):873–81.
20. Gu YU, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific Language model pretraining for biomedical natural Language processing. ACM Trans Comput Healthcare. 2021;3(2):1–23.
21. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art. 2023;6(9):9.
22. Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, et al. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? 2022 ArXiv preprint arXiv:220212837.
23. Zamfrescu-Pereira JD, Hartmann B, Wong R, Yang Q. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. Proc Conf Hum Factors Comput Sys. 2023;1–21.
24. Singhal K, Azizi S, Tu T, Sara Mahdavi S, Wei J, Won Chung H, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80.
25. Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: An Embodied Multimodal Language Model. 2023 ArXiv preprint arXiv:230303378.
26. Huang S, Dong L, Wang W, Hao Y, Singhal S, Ma S, et al. Language Is Not All You Need: Aligning Perception with Language Models. 2023 ArXiv preprint arXiv:230214045.
27. Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: position paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017;58(4):512–21.
28. Beghi E, Sander JW. The ILAE classification of seizures and epilepsies: implications for the clinic. Expert Rev Neurother. 2018;18(3):179–83.
29. Pitkänen A, Löscher W, Vezzani A, Becker AJ, Simonato M, Lukasiuk K, et al. Advances in the development of biomarkers for epilepsy. Lancet Neurol. 2016;15(8):843–56.
30. Thijs RD, Surges R, O'Brien TJ, Sander JW. Epilepsy in adults. Lancet. 2019;393(10172):689–701.
31. Van Donselaar CA, Stroink H, Arts WF. How confident are we of the diagnosis of epilepsy? Epilepsia. 2006;47 Suppl 1(SUPPL. 1):9–13.
32. Savova G, Pestian J, Connolly B, Miller T, Ni Y, Dexheimer JW, et al. Natural Language processing: applications in pediatric research. Translational Bioinformatics. 2016;10:231–50.
33. Buchlak QD, Esmaili N, Bennett C, Farrokhi F. Natural Language processing applications in the clinical neurosciences: a machine learning augmented systematic review. Acta Neurochir Suppl. 2022;134:277–89.
34. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language processing in radiology: a systematic review. Radiology. 2016;279(2):329–43.
35. Cavalleri GL, Petrovski S, Fitzsimons M, Delanty N. eHealth as a facilitator of precision medicine in epilepsy. Biomed Hub. 2017;2(1):137–45.
36. Decker BM, Hill CE, Baldassano SN, Khankhanian P. Can antiepileptic efficacy and epilepsy variables be studied from

electronic health records? A review of current approaches. Seizure. 2021;85:138–44.

37. Connolly B, Matykiewicz P, Cohen KB, Standridge SM, Glauser TA, Dlugos DJ, et al. Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals. J Am Med Inform Assoc. 2014;21(5):866–70.
38. Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford DV, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open. 2019;9(4):e023232.
39. Decker BM, Turco A, Xu J, Terman SW, Kosaraju N, Jamil A, et al. Development of a natural language processing algorithm to extract seizure types and frequencies from the electronic health record. Seizure. 2022;101:48–51.
40. Xie K, Gallagher RS, Shinohara RT, Xie SX, Hill CE, Conrad EC, et al. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. Epilepsia. 2023;64(7):1900–9.
41. Beaulieu-Jones BK, Villamar MF, Scordis P, Bartmann AP, Ali W, Wissel BD, et al. Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study. Lancet Digit Health. 2023;5(12):e882–e894.
42. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. 2022 arXiv preprint arXiv:220111838.
43. Mbizvo GK, Buchan I. Predicting seizure recurrence from medical records using large language models. Lancet Digit Health. 2023;5(12):e851–e852.
44. Kozlowski D, Hutchinson M, Hurley J, Rowley J, Sutherland J. The role of emotion in clinical decision making: an integrative literature review. BMC Med Educ. 2017;17(1):255.
45. Pevy N, Christensen H, Walker T, Reuber M. Feasibility of using an automated analysis of formulation effort in patients' spoken seizure descriptions in the differential diagnosis of epileptic and nonepileptic seizures. Seizure. 2021;91:141–5.
46. Walker LE, Mirza N, Yip VLM, Marson AG, Pirmohamed M. Personalized medicine approaches in epilepsy. J Intern Med. 2015;277(2):218–34.
47. Hakeem H, Feng W, Chen Z, Choong J, Brodie MJ, Fong SL, et al. Development and validation of a deep learning model for predicting treatment response in patients with newly diagnosed epilepsy. JAMA Neurol. 2022;79(10):986–96.
48. Cronin W, Kwan P, Foster E. Anxiety and depressive symptoms in adults with new-onset seizures: a scoping review. Epilepsia Open. 2023;8:758–72.
49. Smolyansky ED, Hakeem H, Ge Z, Chen Z, Kwan P. Machine learning models for decision support in epilepsy management: a critical review. Epilepsy Behav. 2021;123:108273.
50. Vulpius SA, Werge S, Jørgensen IF, Siggaard T, Biel JH, Knudsen GM, et al. Text mining of electronic health records can validate a register-based diagnosis of epilepsy and subgroup into focal and generalized epilepsy. Epilepsia. 2023;64(10):2750–60.
51. Kraljevic Z, Shek A, Bendayan R. Foresight-Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs. 2022 ArXiv preprint arXiv:221208072.
52. Wu X, Zhang T, Zhang L, Qiao L. Epileptic seizure prediction using successive variational mode decomposition and transformers deep learning network. Front Neurosci. 2022;16:982541.
53. Kim H-W, Shin D-H, Kim J, Lee G-H, Cho JW. Assessing the performance of ChatGPT's responses to questions related to epilepsy: A cross-sectional study on natural language processing and medical information retrieval. Seizure. 2024;114:1–8.
54. Davis KL, Candrilli SD, Edin HM. Prevalence and cost of nonadherence with antiepileptic drugs in an adult managed care population. Epilepsia. 2008;49(3):446–54.
55. Faught E, Duh MS, Weiner JR, Guérin A, Cunnington MC. Nonadherence to antiepileptic drugs and increased mortality. Neurology. 2008;71(20):1572–8.
56. Al-aqeel S, Gershuni O, Al-sabhan J, Hiligsmann M. Strategies for improving adherence to antiepileptic drug treatment in people with epilepsy. Cochrane Database Syst Rev. 2020;2(10):CD008312.
57. Sepat R, Sinha AP, Murry LL, Parihar J, Singh MB, Pandey S. Does an additional virtual interactive session increase the impact of digital educational material given to epilepsy patients? A randomized controlled trial. Seizure. 2021;92:252–6.
58. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. Physicians' working hours and lowers their career satisfaction. Int J Health Serv. 2014;44(4):635–42.
59. Dratsch T, Chen X, Mehrizi MR, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. 2023;307(4):e222176.
60. Dahmen J, Kayaalp ME, Ollivier M, Pareek A, Hirschmann MT, Karlsson J, et al. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. Knee Surg Sports Traumatol Arthrosc. 2023;31(4):1187–9.
61. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural Language generation. ACM Comput Surv. 2023;55(12):1–38.
62. Nori H, King N, Mckinney SM, Carignan D, Horvitz E. Openai M 2. Capabilities of GPT-4 on Medical Challenge Problems. 2023 ArXiv preprint arXiv:230313375.
63. Jiang K, Nordli DR, Galan F. The devil is in the details: understanding how misinformation regarding epilepsy manifests in TikTok videos. Epileptic Disord. 2023;25(1):28–32.
64. Venkit P, Srinath M, Wilson S. A study of implicit bias in pretrained language models against people with disabilities. Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea: International Committee on Computational Linguistics; 2022. p. 1324–32.
65. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang P-S, et al. Ethical and social risks of harm from Language Models. 2021 ArXiv preprint arXiv:211204359.
66. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Proces Syst. 2020;33:1877–901.
67. Bano M, Zowghi D, Gervasi V, Shams R. AI for All: Operationalising Diversity and Inclusion Requirements for AI Systems. 2023 ArXiv preprint arXiv:231114695.
68. Brigo F, Otte WM, Igwe SC, Tezzon F, Nardone R. Brief communication clearly written, easily comprehended? The readability of websites providing information on epilepsy. Epilepsy Behav. 2015;44:35–9.
69. Hagendorff T, Fabi S. Methodological reflections for AI alignment research using human feedback. 2022 ArXiv preprint arXiv:230106859.

70. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models. 2022 ArXiv preprint arXiv:220607682.

71. Amodei D, Olah C, Brain G, Steinhardt J, Christiano P, Schulman J, et al. Concrete Problems in AI Safety. 2016 ArXiv preprint arXiv:160606565.

72. Gottlieb S, Silvis L. How to safely integrate large Language models into health care. JAMA Health Forum. 2023;4(9):e233909.

73. Huang P-S, Zhang H, Jiang R, Stanforth R, Welbl J, Rae JW, et al. Findings of the Association for Computational Linguistics Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. 2019 ArXiv preprint arXiv:191103064.

74. Carlini N, Jagielski M, Papernot N, Terzis A, Tramer F, Zhang C, et al. The privacy onion effect: memorization is relative. Adv Neural Inf Proces Syst. 2022;35:13263–76.

75. Karabacak M, Margetis K. Embracing large Language models for medical applications: opportunities and challenges. Cureus. 2023;15(5):e39305.

76. Winata GI, Madotto A, Lin Z, Liu R, Yosinski J, Fung P. Language Models are Few-shot Multilingual Learners. 2021 ArXiv preprint arXiv:210907684.

77. Muennighoff N, Wang T, Sutawika L, Roberts A, Biderman S, Le Scao T, et al. Crosslingual Generalization through Multitask Finetuning. 2022 ArXiv preprint arXiv:221101786.

78. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. 2022 ArXiv preprint arXiv:220405862.

79. Köpf A, Kilcher Y, Von Rütte D, Anagnostidis S, Tam Z-R, Stevens K, et al. OpenAssistant Conversations-Democratizing Large Language Model Alignment. 2023 ArXiv preprint arXiv:230407327.

80. Ganapini MB, Campbell M, Fabiano F, Horesh L, Lenchner J, Loreggia A, et al. Combining Fast and Slow Thinking for Human-like and Efficient Navigation in Constrained Environments. 2022 ArXiv preprint arXiv:220107050.

81. Yue Zhuo T, Huang Y, Chen C, Xing Z. Red teaming ChatGPT via jailbreaking: bias, robustness, reliability and toxicity. 2023 ArXiv preprint arXiv:230112867.

82. Liu Y, Deng G, Xu Z, Li Y, Zheng Y, Zhang Y, et al. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. 2023 ArXiv preprint arXiv:230513860.

83. Li H, Guo D, Fan W, Xu M, Huang J, Meng F, et al. Multi-step Jailbreaking Privacy Attacks on ChatGPT. 2023 ArXiv preprint arXiv:230405197.

84. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med. 2023;29:2396–8.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: van Diessen E, van Amerongen RA, Zijlmans M, Otte WM. Potential merits and flaws of large language models in epilepsy care: A critical review. Epilepsia. 2024;65:873–886. https://doi.org/10.1111/epi.17907