BRITISH JOURNAL OF HOSPITAL MEDICINE

REVIEW

Large Language Models: A Paradigm Shift for Dementia Diagnosis and Care

Judith Rose Harrison[1,*], Song Ling Tang[2], Bede Burston[3], Alexander Robertson[4], Huizhi Liang[4], John Paul Taylor[1]

[1]Translational and Clinical Research Institute, Newcastle University, Newcastle, UK
[2]Hertfordshire Partnership University NHS Foundation Trust, St Albans, UK
[3]School of Medicine, Newcastle University, Newcastle, UK
[4]School of Computing, Newcastle University, Newcastle, UK
*Correspondence: Judith.harrison@newcastle.ac.uk (Judith Rose Harrison)

Abstract

Dementia poses major challenges to healthcare worldwide. Traditional diagnostics rely on lengthy assessments, and access to specialist clinicians is limited. Large language models (LLMs), like Generative Pre-trained Transformer 4 (GPT-4) present new avenues for enhancing dementia diagnosis and care through advanced language processing. Whilst research into their applications is in its infancy, LLMs can harness vast datasets and powerful algorithms, with significant potential to enhance diagnostic accuracy in dementia, monitor symptom progression, and provide personalised care recommendations. While dementia serves as the primary example, the ethical and practical considerations discussed are applicable to the wider use of LLMs across different areas of medicine. This review explores the prospect of LLMs transforming dementia management and addresses the ethical and practical considerations involved.

Key words: dementia; large language models; diagnosis; patient care; artificial intelligence

Submitted: 19 September 2024 Revised: 23 January 2025 Accepted: 4 February 2025

Introduction

Neurodegenerative dementias cause progressive cognitive decline. This has a devastating effect on both those diagnosed and their families. With an ageing global population, neurodegenerative diseases have become an urgent healthcare challenge. Over 55 million individuals are currently living with dementia worldwide and this is expected to increase to 115.4 million in 2050 (Prince et al, 2013). The development of new amyloid-targeting drugs for Alzheimer's Disease adds to the strain on diagnostic services, as they necessitate precise and early detection of amyloid pathology for disease modification (Koychev et al, 2024). Additionally, neurocognitive diagnostic assessments are often inaccurate, with approximately one-quarter of patients receiving a different dementia diagnosis at autopsy (Gauthreaux et al, 2020). Innovative solutions are needed to improve dementia diagnosis, treatment, and caregiving, reducing the global burden of the disease.

Advanced artificial intelligence (AI) tools, such as large language models (LLMs), offer significant promise in healthcare (Min et al, 2023). LLMs are an advanced form of natural language processing (NLP). Unlike early rule-based NLP systems,

How to cite this article:
Harrison JR, Tang SL, Burston B, Robertson A, Liang H, Taylor JP. Large Language Models: A Paradigm Shift for Dementia Diagnosis and Care. Br J Hosp Med. 2025. https://doi.org/10.12968/hmed.2024.0666

modern LLMs have greatly evolved (see **Supplementary Fig. 1**, for details). Developed using advanced neural networks and transformers (Hughey and Krogh, 1996), LLMs are capable of understanding and generating natural language. They are also capable of processing vast amounts of medical data. This capability is being harnessed to improve memory assessment and intervention by analysing clinical records, speech patterns, and neuroimaging data. It could be transformative for diagnostics, treatments, and patient-caregiver communication in dementia care and research (Meng et al, 2024). However, they also present ethical, practical, and scientific challenges (Li et al, 2023). This review examines the benefits and risks of using LLMs in diagnosis and treatment, using dementia as an example, aiming for their informed and careful integration into practice.

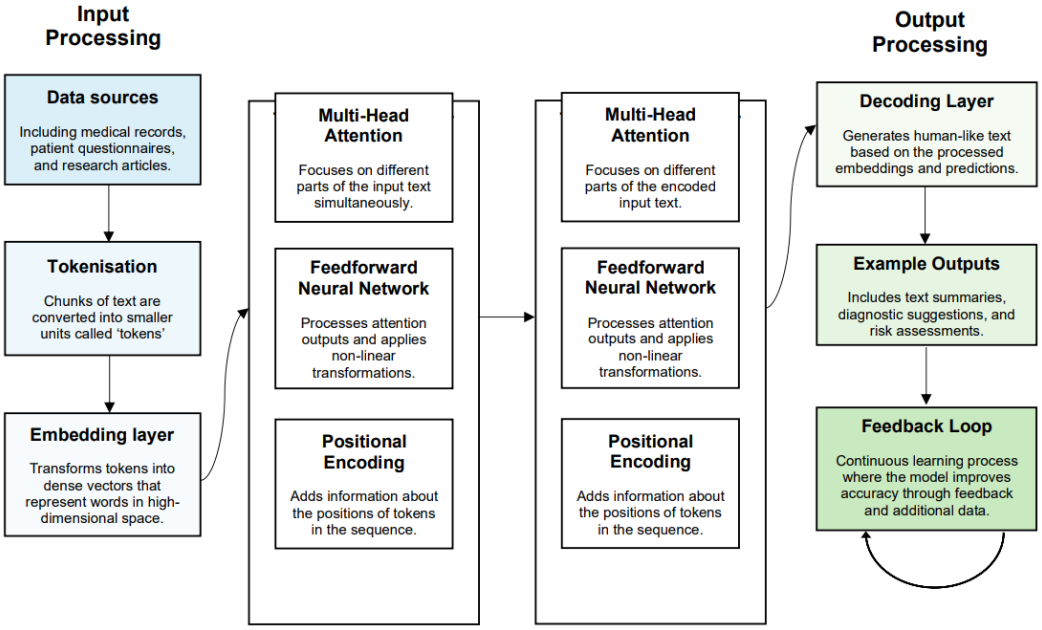

Fig. 1. Architecture and functioning of large language models (LLMs). This figure illustrates the steps involved in processing text data using a transformer-based LLM architecture. The process begins with Input Processing, where diverse data sources, such as medical records and patient questionnaires, are tokenised and embedded. The Transformer Encoder and Decoder components apply multi-head attention and positional encoding to understand the context and relationships within the text, supporting the generation of human-like output. In Output Processing, the model produces example outputs like diagnostic suggestions and risk assessments, which may aid clinicians in evidence-based decision-making. The Feedback Loop enables continuous learning, improving model accuracy over time with additional data. Software used for figure creation: Microsoft PowerPoint (Stable release 2013; Microsoft, Redmond, WA, USA).

Fig. 1 shows transformer-based LLM architecture. Examples of LLMs include Google's bidirectional encoder representations from transformers (BERT) (Devlin et al, 2019) and Chat Generative Pre-trained Transformer (ChatGPT) (Radford et al, 2019). Several companies have developed open-source LLMs, which could be adapted for academic or clinical use (See **Supplementary Material** for examples of these). The process of LLM training, shown in Table 1, demonstrates how these

Table 1. Stages of training large language models (LLMs).

Stage	Pre-training	Fine-tuning	In-context learning
Objective	Teach general linguistic and healthcare patterns	Specialise the model in dementia-specific domains	Adapt the model to patient-specific needs during interaction
Method	Train on a large corpus, including healthcare texts	Continue training on dementia datasets and clinical guidelines	Use patient-specific prompts, such as 'Provide caregiver information'
Outcome	Encodes general medical knowledge	Enhances dementia-specific diagnostic and care capabilities	Provides personalised and contextually relevant responses
Type	Adjusting model weights	Adjusting model weights	Enhancing patient interactions without altering model structure
Computational Intensity	High	High	Low
Relevance to neurodegenerative conditions	Learns general dementia concepts, such as symptoms and terminology	Specialises in diagnosing and managing dementia subtypes	Adapts responses to support users (e.g., caregivers or clinicians) with real-time feedback
Key references	(Vaswani et al, 2017)	(Min et al, 2023)	(Wei et al, 2022; Yao et al, 2024; Yasunaga et al, 2024)

models could be tailored to improve diagnostic accuracy and management of patients with neurodegenerative conditions.

After training, LLMs can incorporate additional knowledge using techniques like Retrieval-Augmented Generation (RAG), which integrates external, up-to-date data. This approach could be particularly useful for evidence-based recommendations and clinical decision support in memory assessment and therapeutics. RAG could be used to improve diagnostic accuracy by combining real-time patient data with clinical knowledge. However, to ensure relevant and coherent responses in dementia contexts, high-quality data and seamless integration are essential, and validation in neurocognitive clinical settings is needed (Chen et al, 2024; Gao et al, 2024).

Machine Learning in Dementia

Machine learning (ML) methods are established in dementia research. ML includes techniques like Support Vector Machines (SVMs), random forests and convolutional neural networks (CNNs). For instance, various ML algorithms have been used to analyse neuroimaging data to detect early signs of Alzheimer's Disease, with deep learning models achieving high accuracy in identifying Alzheimer's years before clinical symptoms appear (Borchert et al, 2023). ML can also enhance insights from cognitive testing data, potentially improving the accuracy and scalability of neurocognitive screening tools (Li et al, 2022). ML can identify speech and linguistic markers associated with cognitive decline (Brewer et al, 2021) and electroencephalogram (EEG) features that could serve as biomarkers for diagnosis and disease progression (Jiao et al, 2023).

ML with multi-modal data could improve dementia diagnostic accuracy. For example, a recent study tested a multi-modal ML model designed to perform differential diagnosis of dementia using a variety of data including demographics, medical history, neuroimaging, and neuropsychological assessments. The model was validated on over 51,000 participants from diverse datasets and demonstrated high accuracy in distinguishing between normal cognition, Mild Cognitive Impairment (MCI), and dementia, as well as different dementia etiologies. The model outperformed clinician-only assessments, aligning with biomarker and postmortem findings, suggesting its potential for integration into clinical settings to enhance memory screening and management (Xue et al, 2024). An example of a ML model predicting progression from Mild Cognitive Impairment to Alzheimer's Disease is summarised in Box 1 (Lin et al, 2020). However, none of these ML methods has yet crossed the translational divide into routine clinical care.

Emerging Applications of LLMs in Dementia Research and Clinical Care

While traditional ML models have shown great promise, particularly in the diagnosis of neurodegenerative diseases, LLMs are emerging as powerful complementary tools (Zheng et al, 2025). Their abilities go far beyond processing and generating human language. LLM techniques can be combined to allow integra-

tion of narrative information, such as the history provided by a patients' caregivers, alongside biomarkers, including neuroimaging using embedded neural networks and ML algorithms (Feng et al, 2023). Although the field of LLMs in dementia is still nascent, the integration of these models with established ML methods will significantly enhance both research and clinical practice (Meng et al, 2024).

This case study demonstrates an ML model, specifically an Extreme Learning Machine (ELM), which combines multiple types of data to predict Mild Cognitive Impairment to Alzheimer's Disease (MCI-to-AD) conversion with enhanced accuracy.

Data Sources and Processing

- **MRI**: Provides detailed images of brain structure, allowing for the assessment of atrophy patterns in areas commonly affected by AD.
- **FDG-PET**: Measures glucose metabolism in the brain, with reduced metabolic activity often observed in AD-affected regions.
- **CSF Biomarkers**: Levels of amyloid-beta and tau proteins are early indicators of AD pathology and are increasingly measured in clinical practice.
- **Genetics**: Information on APOE ϵ4 status, a genetic marker associated with higher AD risk, adds a hereditary risk factor to the analysis.

Machine Learning Model – Extreme Learning Machine (ELM)

- Each data type (MRI, FDG-PET, CSF, and genetics) was first analysed individually to produce scores that indicate how closely an MCI patient's profile resembles those of AD or control groups.
- These scores were then combined and processed through the ELM model, which produced a final prediction on whether the patient's MCI would likely progress to AD.

Results

- This multimodal ML approach achieved an 84.7% accuracy in predicting MCI-to-AD conversion over a 3-year period.
- When compared to using single data sources, integrating multiple data types improved prediction accuracy by about 10%.

Box 1. Case Study—The Extreme Learning Machine: Using Machine Learning to Predict Alzheimer's Disease Conversion from Mild Cognitive Impairment (Lin et al, 2020). Acronyms: ML, Machine Learning; AD, Alzheimer's Disease; MRI, Magnetic Resonance Imaging; FDG-PET, Fluorodeoxyglucose Positron Emission Tomography; CSF, Cerebrospinal Fluid; APOE, Apolipoprotein E.

Models based on Google's BERT are established in genomics, methylomics and transcriptomics research (Liu et al, 2024). For example, a version of BERT trained on genomic data—known as DNABERT—has shown state-of-the-art performance in predicting which segments of the genome are regulatory elements (Ji et al, 2021). This will improve the annotation and interpretation of genomic data, facilitating the identification of genetic risk factors for dementia (Ji et al, 2021). There are hopes that the application of LLMs in molecular biology, trained on large

proteineamic datasets, could enable researchers to pinpoint novel drug targets for neurodegenerative conditions (Tripathi et al, 2024). This is a major area of investment in the pharmaceutical industry at present (Arnold, 2023).

LLMs can streamline the analysis of scientific literature, aiding systematic reviews in dementia. For example, a recent study tested a number of open-source LLMs with different prompt designs to screen titles and abstracts based on predefined inclusion criteria (Dennstädt et al, 2024). They reported that some classifiers and models achieved high sensitivity and specificity, but noted that such techniques remain exploratory and should not be relied on (Dennstädt et al, 2024).

Recruitment of memory clinic patients to clinical trials could be streamlined using LLMs. In a detailed experiment, Yuan et al (2024) tested a version of BERT for improving patient-trial matching in healthcare. They used data from the electronic health records (EHR) of stroke patients, and from Medical Information Mart for Intensive Care (MIMIC-III), a well-known critical care database, alongside clinical trial criteria from stroke studies. Their proposed pipeline, LLM-based Patient Trial Matching (LLM-PTM), used privacy-conscious methods to augment the language model with patient data. The enriched semantic knowledge produced a 7.32% improvement in patient-trial matching and a 12.12% increase in generalisability (Yuan et al, 2024). Such techniques could be particularly useful for expedited recruitment to trials of new dementia therapies. Similar methods could be applied to facilitate EHR-based studies focused on understanding disease progression and evaluating the efficacy of treatments (Wang et al, 2023).

Clinical applications of LLMs in neurocognitive disorders are still in the experimental phase. Speech analysis for detecting cognitive impairment is the most advanced application of LLMs, with several studies using actual patient data (Agbavor and Liang, 2022; Bang et al, 2024; Cohen and Pakhomov, 2020; Guo et al, 2019). Box 2 shows a Case Study of this. This approach has shown promise in identifying early signs of cognitive decline through the analysis of language patterns, and distinguishing Alzheimer's Disease patients from healthy controls (Bang et al, 2024). It is a potential non-invasive LLM tool for detecting dementia.

Several groups have conducted proof-of-concept studies using simulated clinical data. LLM tools aiming to provide cognitive support, enhance social interaction, and offer companionship for individuals with dementia are in development (Favela et al, 2023; Hossain et al, 2024). One study used simulated patients to investigate the potential to automate cognitive stimulation therapy for neurodegenerative diseases (Favela et al, 2023). Others have assessed the quality of LLM responses to dementia caregivers' questions, using either simulated scenarios and dialogues, or questions taken from social media (Aguirre et al, 2024; Saeidnia et al, 2024). Generally, the LLM provided high-quality information that was relevant to the scenario, although it lacked knowledge about the patient's context (Aguirre et al, 2024; Saeidnia et al, 2024). This is a significant limitation, as contextual information, such as the structure of local health and social care services, is particularly important in dementia care.

This study tests the effectiveness of using a large language model, GPT-3, to assess dementia through spontaneous speech. Language impairment, an early indicator of Alzheimer's Disease (AD), manifests in speech patterns, vocabulary use, and sentence structure. By analysing these characteristics, GPT-3 can assist in early diagnosis without the need for invasive tests.

Methodology

1. **Speech-to-Text Conversion**: Audio recordings of patient speech were converted to text using Wav2Vec 2.0, a model designed for accurate speech transcription.
2. **Text Embeddings with GPT-3**: GPT-3 generated text embeddings (vectorised representations) from the transcribed speech. These embeddings capture linguistic nuances, such as word choice, syntax, and semantics, which are relevant for detecting cognitive decline.
3. **Classification and Prediction**: Machine learning models, including support vector classifiers and random forests, were trained using the GPT-3 embeddings. These models performed two key tasks:
    - **AD Classification**: Differentiating between individuals with AD and healthy controls based solely on speech data.
    - **Mini-Mental State Examination (MMSE) Score Prediction**: Predicting cognitive test scores from speech features.

Results

- GPT-3 embeddings significantly outperformed traditional acoustic feature-based methods, achieving high accuracy in both AD classification and MMSE score prediction.
- Notably, the GPT-3-based approach was competitive with fine-tuned models, demonstrating the potential for LLMs to serve as effective tools in dementia screening.

Box 2. Case Study—Predicting Dementia from Spontaneous Speech Using GPT-3 (Agbavor and Liang, 2022). Acronyms: GPT-3, Generative Pre-trained Transformer 3; LLM, Large Language Model.

LLMs are being studied as potential clinical decision-making tools for diagnosing neurocognitive diseases. The LLM was presented with clinical vingettes including descriptions of their patients' complaints, physical examination, as well as biomarkers, neuroimaging findings, and neuropsychological data, and was asked to give the likely dementia subtype diagnosis. The results demonstrated clear concordance with human raters, albeit in a very small sample of synthetic cases (El Haj et al, 2024).

The attention mechanism used in many LLMs is particularly well-suited for analysing multi-modal data. A recent study applied ChatGPT to the problem of Alzheimer's Disease diagnosis (Feng et al, 2023). Feng et al (2023) used publicly available data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, which includes clinical, neuropsychological, genetic, and imaging data.

Table 2. Comparison of traditional methods vs. potential LLM-enhanced clinical methods for neurocognitive disease assessment and management.

Aspect	Traditional methods	LLM-enhanced methods	Reference
Clinical data collection	Manual data entry from patient visits	Ambient voice recording and wearable devices, integrated with Electronic Health Records (EHR)	(Topol, 2019)
Data processing	Typically based on clinician judgment	Advanced natural language processing (NLP) and machine learning algorithms	(Jiang et al, 2021)
Risk identification	Typically based on clinician judgment	EHR data-driven risk factor identification	(Andrews et al, 2024)
Symptom tracking & Remote monitoring	Periodic assessments	Continuous monitoring via wearable devices and questionnaires	(Knapp et al, 2022; Xie et al, 2020)
Decision support	Clinical guidelines and experience	Real-time recommendations based on comprehensive data	(Meng et al, 2024)
Care plans	Standardised care plans	Tailored care plans using patient-specific data	(Knapp et al, 2022)
Continuous learning	Limited updates based on new research	Continuous improvement through ongoing data training	(Meng et al, 2024)

Various methods, including embedding and modality alignment, were used to process and fuse data types like Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and non-image information. The LLM-based model achieved state-of-the-art performance, surpassing traditional ML models in correctly classifying participants as Alzheimer's Disease or MCI patients, or normal controls (Feng et al, 2023). This study included data from over 400 research participants, rather than simulated patients. However, it should be noted that the highly phenotyped subjects included in the ADNI cohort may not be representative of dementia cases in other populations. Further replication of these methods is necessary.

There are ongoing investigations into the effectiveness of human-AI collaboration in healthcare, particularly in medical diagnosis. While AI can enhance diagnostic accuracy and efficiency (Cabitza et al, 2023), human expertise remains crucial for interpreting contextual information and maintaining ethical standards (Formosa et al, 2022; Sezgin, 2023). Research suggests that human-AI teams can outperform both humans and AI alone in various medical tasks, including radiology (Cacciamani et al, 2023) and endoscopy (Reverberi et al, 2022). However, the success of human-AI collaboration depends on factors such as AI interpretability, predictability, and control (Bienefeld et al, 2024). Examples of potential LLM technology for cognitive evaluation and therapeutics are shown in Table 2.

Whilst the potential of LLMs in research and clinical care is clear, there are a number of knowledge and capability gaps. For example, LLMs may struggle to interpret imperfect real-world multi-modal data. It may prove difficult to integrate LLMs with EHRs or to provide appropriate contextual information. Further replication of exploratory studies is required, including empirical testing on actual clinical data. Although much of the available literature on LLMs is disseminated

Table 3. Causes and mitigations of hallucinations in LLMs.

Causes of hallucinations in LLMs	Mitigation strategies for hallucinations
Inaccurate or inconsistent training data Errors or inconsistencies in dementia datasets can lead to incorrect suggestions around diagnosis or management (Ye et al, 2024; Zhang et al, 2023).	**High-quality data** Using well-curated dementia-specific training data and allowing users to flag errors improves accuracy over time (Liu et al, 2023).
Model constraints LLM interpretative limitations can result in incorrect conclusions, especially for subtle cognitive symptoms (Ye et al, 2024; Zhang et al, 2023).	**Explainability** Requiring LLMs to provide references for their claims allows clinicians to verify the source of information, reducing the risk of hallucinations (Liu et al, 2023).
Prompt quality Low-quality or ambiguous prompts can lead to hallucinations, such as misidentifying dementia subtypes. (Zhang et al, 2023).	**Self-verification** Fact-checking tools or chain-of-thought reasoning, where the model generates intermediate steps to improve problem-solving accuracy, help ensure the accuracy of outputs against the model's knowledge base (Dhuliawala et al, 2024; Wei et al, 2022).
Poor calibration Poor calibration, or the lack of adjustment in a model's confidence levels, can cause LLMs to overestimate their certainty in factually incorrect dementia-related outputs, leading to potentially misleading conclusions (Jiang et al, 2021).	**Continuous monitoring** Metrics like TruthfulQA help quantify and monitor hallucinations, enabling continuous improvement in dementia-related applications (Lin et al, 2022).

through pre-print servers like arXiv, clinical applications—especially for serious health conditions like dementias—must undergo rigorous peer-reviewed validation. Finally, although initial research indicates that these technologies are generally well-received by patients and caregivers (Treder et al, 2024), concerns persist regarding their acceptability, particularly if their use reduces face-to-face interactions with clinicians.

When LLMs Go Wrong—And What to Do About It

Hallucinations

LLMs function as completion engines, predicting the next word based on their training data. In LLMs, 'hallucinations' describe generated text that sounds convincing but is factually incorrect (IBM, 2023). This might include producing incorrect treatment suggestions that could mislead clinicians, patients or caregivers, potentially leading to harmful interventions. There are several reasons why this can happen, and strageies have been developed to address them (See Table 3).

Bias

LLMs are known to be susceptible to bias (Obradovich et al, 2024). Studies have shown that GPT-4 exhibits encoded racial and gender biases, raising serious concerns about the potential harm these biases could cause when using LLMs for

medical decision support (Zack et al, 2024). Furthermore, biases in LLMs could disproportionately affect vulnerable populations, such as older adults, by reinforcing stereotypes or underrepresenting specific ethnic groups in neurocognitive disease diagnosis (see Case Study in Box 3). Mitigation strategies for bias, tailored to dementia-focused LLMs, should involve rigorous validation against high-quality, dementia-specific datasets, regular bias audits (systematic evaluations conducted to detect and measure biases in outputs), and continual monitoring to ensure clinical safety, as described in Table 4.

This study examines racial biases in large language models (LLMs), specifically GPT-3.5-turbo and GPT-4, when generating medical reports. Using synthetic patient profiles, researchers assessed how LLM outputs varied across racial and ethnic groups, revealing potential biases in treatment recommendations, cost projections, and survival predictions.

Methodology

1. **Patient Profile Creation**: Synthetic profiles representing patient groups racialised as white, Black, Hispanic, and Asian were created using de-identified data from PubMed Central. Profiles included variations in demographics, medical history, and racial identifiers to test for differential outputs.
2. **Prompting LLMs for Reports**: LLMs were prompted to generate medical reports for each profile, providing predictions on hospitalisation duration, treatment recommendations, projected costs, and survival outcomes. Each profile's race was systematically altered to isolate the effect of racial identifiers on LLM outputs. al stay, and associated costs across racial groups.
3. **Comparative Analysis**: The generated reports were analyzed to identify differences in treatment aggressiveness, length of hospital stay, and associated costs across racial groups.

Results

- LLMs projected longer hospital stays, higher costs, and more aggressive treatment recommendations for white patients compared to other groups.
- For Black patients, LLMs tended to suggest more conservative treatment approaches, such as ICU care over surgical intervention, even when medical profiles were otherwise identical.

Box 3. Case Study—Racial Bias in LLM-Generated Medical Reports (Yang et al, 2024).

Ethical and Regulatory Considerations in AI-Driven Dementia Care: Privacy, Transparency, Fairness, and Accountability

Privacy

Using generative AI models, especially when trained on personal health data, raises significant privacy concerns, particularly in sensitive fields like the treatment

Table 4. Causes and mitigations of bias in LLMs.

Causes of bias in LLMs	Mitigation strategies for bias
Bias in training data LLMs can learn and perpetuate biases from unbalanced or biased datasets, reinforcing stereotypes and underrepresenting certain groups in dementia (Blodgett et al, 2020).	**Pre-processing** Ensure diverse and balanced training data to reduce biases, such as ensuring equal representation of different demographic groups in dementia studies, with appropriate data annotation (Bender et al, 2021).
Bias in learning processes The model's learning process itself may amplify existing biases, leading to unfair treatment or recommendations for certain demographic groups (Treder et al, 2024).	**In-training adjustments** Modify the training process by adapting model architecture to minimise bias during learning (Lauscher et al, 2021).
Encoded biases Pre-existing racial or gender biases in LLMs like GPT-4 can affect medical decision support, raising concerns for dementia diagnosis and treatment (Zack et al, 2024).	**Intra-processing** Apply bias reduction techniques post-training to filter or rewrite biased outputs, improving fairness in dementia-related responses (Gallegos et al, 2024).
	Post-processing Rewriting or filtering generated text post-training can help mitigate biased suggestions or conclusions (Gallegos et al, 2024).
	Self-correction Use chain-of-thought reasoning and model instructions to guide LLMs towards more unbiased outputs, particularly for sensitive medical contexts like dementia diagnostic assessments (Ganguli et al, 2023; Wei et al, 2022).

of older adults (Ong et al, 2024). Patients with cognitive impairments may unknowingly contribute personal data without informed consent, heightening ethical challenges. This is essential in clinical research, where stringent data protection is required. Compliance with regulations such as the General Data Protection Regulation (GDPR) and the UK Data Protection Act 2018 is necessary to protect vulnerable patients and maintain trust in AI-driven dementia diagnosis and care tools (Health Research Authority, 2018; Meskó and Topol, 2023; NHS England, 2023).

To further strengthen patient privacy, advanced anonymisation techniques such as differential privacy should be considered. Even anonymised data can sometimes be re-identified, particularly in cases where it is combined with other spatiotemporal data points, increasing risks of privacy breaches (Meskó and Topol, 2023). Additional cybersecurity measures, including regular penetration testing and resilience against adversarial attacks, are essential to protect sensitive health data from unauthorised access and adversarial attacks. Rigorous benchmarks to evaluate the balance between privacy and model utility are also necessary, ensuring that LLMs used in dementia care deliver valuable clinical insights while safeguarding confidentiality (He et al, 2023).

Transparency, Explainability, and Data Provenance

Transparency in data collection and AI decision-making is essential for clinical use, especially in cognitive evaluation, where patient outcomes rely on accurate and explainable models. LLMs are often seen as 'black boxes' due to their complex algorithms, making it challenging to understand their decision-making process (Meskó and Topol, 2023). Companies like Google, Microsoft, and OpenAI, which have produced many of the open-source LLMs suitable for dementia innovations, often limit external scrutiny, hindering verification of their capabilities. This lack of transparency complicates accountability in AI-driven dementia diagnoses (NHS England, 2023). Addressing this issue also requires education for clinicians on AI fundamentals. Limited AI literacy among clinical educators, combined with competing demands on curricula, could perpetuate the current skills gap. Closing this gap is essential to prepare clinicians to interpret and responsibly apply AI outputs (Misra et al, 2024).

To build trust, developers should prioritise transparent architectures and explainable AI techniques. These approaches clarify how models arrive at predictions, allowing clinicians to make better-informed decisions based on AI outputs (Mökander et al, 2024). Transparency should also extend to data provenance, as LLMs often ingest data from a wide array of sources, some of which may include proprietary content. Ensuring that training data is legally and ethically sourced minimises the risk of intellectual property violations and supports ethical accountability in clinical applications of LLMs (Minssen et al, 2023).

Regulatory Frameworks for AI-Enabled Healthcare

Regulators globally are working to effectively oversee healthcare innovations involving LLMs, including for neurodegenerative diseases. The UK Government has chosen a principles-based framework outlined in its AI White Paper published in March 2023 (Department for Science, Innovation & Technology, 2024). This approach enables existing regulators, such as the Medicines and Healthcare Products Regulatory Agency (MHRA), to apply flexible, context-specific guidance tailored to respective sectors (Dennis and Vollers, 2023; Yaros et al, 2022). In an attempt to identify and address the challenges associated with AI in medical devices, the MHRA has developed an AI and Software as Medical Device (SaMD) Airlock (Medicines and Healthcare Products Regulatory Agency, 2024). This initiative aims to safely integrate AI technologies, including AI-driven medical devices, into healthcare by offering developers a controlled environment to test and refine AI solutions while ensuring they meet safety and quality standards.

To establish fairness, particularly in healthcare, it is important to test LLMs for biases that may disadvantage certain patient groups. As discussed in Table 4, bias audits, representative datasets, and continuous stewardship help reduce inequities and improve the inclusivity of AI models. This is particularly crucial in dementia diagnosis, as LLMs may inherit biases from training data that underrepresents minoritised communities, where dementia is often underdiagnosed or diagnosed later in the UK (Gove et al, 2021; Tsamakis et al, 2021). This disparity in diagnostic timing and frequency could skew predictive models, potentially leading to reduced ac-

curacy or missed early diagnoses for these populations (Farina and Lavazza, 2023). Ongoing algorithmic monitoring and adjustment are essential, as they allow AI models to evolve while maintaining fairness and safety over time (Liu et al, 2022).

Patient Autonomy and the Right to Data Control

Determining who is liable when AI-generated recommendations lead to patient harm is complex, particularly in memory assessment and treatment, where decisions are nuanced. To date, healthcare systems have always required a learned intermediary or a human-in-the-loop. This is especially relevant in dementia diagnosis and management, where AI can assist but should not replace clinical judgment. The General Medical Council (GMC) emphasises that doctors using AI systems are responsible for their clinical decisions, even when influenced by AI-generated recommendations. Ultimately, clinicians must ensure that AI does not compromise the safety of patients in dementia services (General Medical Council, 2024).

Respecting patient autonomy is essential when using personal health data in LLMs. Patients should be aware of their rights, including the right to access, rectify, and delete their data, which is especially pertinent in dementia care given cognitive impairments that might limit patients' understanding of these rights (European Data Protection Supervisor, 2023). Consent procedures should clearly explain the possible secondary uses of health data for future AI model improvement, and patients should have the option to limit data processing. Adhering to these principles not only strengthens ethical compliance but also fosters trust in AI-driven dementia care solutions (Reddy et al, 2020).

Conclusion

LLMs have significant potential to enhance both neurodegenerative disease research and clinical care. Adopting advanced technologies is imperative to meet the demographic and treatment challenges posed by dementia in the coming century. Addressing the ethical, practical, and regulatory challenges is of utmost importance, ensuring data privacy, mitigating biases, and establishing clear guidelines for liability and transparency. Future research should explore integrating multimodal data—such as imaging, genetics, and clinical notes—to create a more holistic understanding of dementia. Additionally, exploring the feasibility of combining LLMs with other ML models may further optimise diagnostic accuracy and personalised patient care. With a critical, informed approach, LLMs could cross the translational divide from research to practical dementia care, ultimately improving outcomes for patients and caregivers.

Key Points

- Dementia is a growing healthcare challenge globally, requiring innovative solutions for diagnosis, treatment and caregiving to alleviate its burden.
- Large language models (LLMs) have promising potential to improve dementia care by transforming diagnostic methods, treatments, and communication.
- Integrating LLMs in dementia care introduces risks, such as hallucinations and biases.
- To address these challenges, strategies such as high-quality training data, promoting explainability, and self-verification techniques could be employed at different stages of LLM training.
- Rigorous regulation of LLM use in healthcare to address data privacy, transparency, and medical liability to uphold ethical standards.

Availability of Data and Materials

All the data of this study are included in this article.

Author Contributions

JRH, SLT, HL, AR, BB and JPT all made substantial contributions to the conception of this article. JRH and SLT wrote the manuscript with contributions from AR and HL for the technical sections. BB produced the graphics. All authors contributed to the important editorial changes of important content in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgement

Not applicable.

Funding

JRH and JPT are supported by the NIHR Newcastle Biomedical Research Centre (BRC).

Conflict of Interest

JRH holds an NIHR Clinical Academic Lecturership, serves as a Clinical Advisor for Akrivia Health and as an Expert Reviewer for Medical Devices for the

Medicines and Health Regulatory Authority (MHRA). Other authors have no relevant interests to declare.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://www.magonlinelibrary.com/doi/suppl/10.12968/hmed.2024.0666.

References

Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digital Health. 2022; 1: e0000168. https://doi.org/10.1371/journal.pdig.0000168

Aguirre A, Hilsabeck R, Smith T, Xie B, He D, Wang Z, et al. Assessing the Quality of ChatGPT Responses to Dementia Caregivers' Questions: Qualitative Analysis. JMIR Aging. 2024; 7: e53019. https://doi.org/10.2196/53019

Andrews SJ, Jonson C, Fulton-Howard B, Renton AE, Yokoyama JS, Yaffe K, et al. The Role of Genomic-Informed Risk Assessments in Predicting Dementia Outcomes. medRxiv. 2024. https://doi.org/10.1101/2024.04.27.24306488 (preprint)

Arnold C. Inside the nascent industry of AI-designed drugs. Nature Medicine. 2023; 29: 1292–1295. https://doi.org/10.1038/s41591-023-02361-0

Bang JU, Han SH, Kang BO. Alzheimer's disease recognition from spontaneous speech using large language models. ETRI Journal. 2024; 46: 96–105. https://doi.org/10.4218/etrij.2023-0356.

Bender EM, Gebru T, McMillan-Major A. Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Virtual Event Canada: ACM. 2021. https://doi.org/10.1145/3442188.3445922

Bienefeld N, Keller E, Grote G. Human-AI Teaming in Critical Care: A Comparative Analysis of Data Scientists' and Clinicians' Perspectives on AI Augmentation and Automation. Journal of Medical Internet Research. 2024; 26: e50130. https://doi.org/10.2196/50130

Blodgett SL, Barocas S, Daumé III H, Wallach H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2020; 1: 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Borchert RJ, Azevedo T, Badhwar A, Bernal J, Betts M, Bruffaerts R, et al. Artificial intelligence for diagnostic and prognostic neuroimaging in dementia: A systematic review. Alzheimer's & Dementia. 2023; 19: 5885–5904. https://doi.org/10.1002/alz.13412

Brewer E, Mirheidari B, O'Malley R, Reuber M, Christensen H, Blackburn DJ. Characterising spoken interactions of healthy ageing adults with CognoSpeak, a web-based cognitive assessment tool. Alzheimer's & Dementia. 2021; 17: e052913. https://doi.org/10.1002/alz.052913

Cabitza F, Campagner A, Ronzio L, Cameli M, Mandoli GE, Pastore MC, et al. Rams, hounds and white boxes: Investigating human-AI collaboration protocols in medical diagnosis. Artificial Intelligence in Medicine. 2023; 138: 102506. https://doi.org/10.1016/j.artmed.2023.102506

Cacciamani GE, Sanford DI, Chu TN, Kaneko M, De Castro Abreu AL, Duddalwar V, et al. Is Artificial Intelligence Replacing Our Radiology Stars? Not Yet! European Urology Open Science. 2023; 48: 14–16. https://doi.org/10.1016/j.euros.2022.09.024

Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. In Proceedings of the AAAI Conference on Artificial Intelligence 2024; 38: 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728

Cohen T, Pakhomov S. A Tale of Two Perplexities: Sensitivity of Neural Language Models to Lexical Retrieval Deficits in Dementia of the Alzheimer's Type. In Proceedings of the 58th Annual Meeting of

the Association for Computational Linguistics, Online: Association for Computational Linguistics (pp. 1946–1957). 2020. https://doi.org/10.18653/v1/2020.acl-main.176

Dennis A, Vollers N. Regulating AI as a medical device in the UK. 2023. Available at: https://www.taylorwessing.com/en/insights-and-events/insights/2023/07/regulating-ai-as-a-medical-device-in-the-uk (Accessed: 6 July 2024).

Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. Systematic Reviews. 2024; 13: 158. https://doi.org/10.1186/s13643-024-02575-4

Department for Science, Innovation & Technology. A pro-innovation approach to AI regulation: government response. 2024. Available at: https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response (Accessed: 6 July 2024).

Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics. 2019; 1: 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A, et al. Chain-of-Verification Reduces Hallucination in Large Language Models. In Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics. 2024; 1: 3563–3578. https://doi.org/10.18653/v1/2024.findings-acl.212

El Haj M, Boutoleau-Bretonnière C, Gallouj K, Wagemann N, Antoine P, Kapogiannis D, et al. ChatGPT as a Diagnostic Aid in Alzheimer's Disease: An Exploratory Study. Journal of Alzheimer's Disease Reports. 2024; 8: 495–500. https://doi.org/10.3233/ADR-230191

European Data Protection Supervisor. Rights of the Individual. 2023. Available at: https://www.edps.europa.eu/data-protection/our-work/subjects/rights-individual_en (Accessed: 11 November 2024).

Farina M, Lavazza A. ChatGPT in society: emerging issues. Frontiers in Artificial Intelligence. 2023; 6: 1130913. https://doi.org/10.3389/frai.2023.1130913

Favela J, Cruz-Sandoval D, Parra MO. 'Conversational Agents for Dementia using Large Language Mod- Els', 2023 Mexican International Conference on Computer Science (ENC) (pp. 1–7). Guanajuato, Guanajuato, Mexico. 11-13 September 2023. IEEE: New York, NY, USA. 2023. https://doi.org/10.1109/ENC60556.2023.10508610

Feng Y, Xu X, Zhuang Y, Zhang M. 'Large Language Models Improve Alzheimer's Disease Diagnosis Using Multi-Modality Data', 2023 IEEE International Conference on Medical Artificial Intelligence (MedAI) (pp. 61–66). Beijing, China. 18-19 November 2023. IEEE: New York, NY, USA. 2023. https://doi.org/10.1109/MedAI59581.2023.00016

Formosa P, Rogers W, Griep Y, Bankins S, Richards D. Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. Computers in Human Behavior. 2022; 133: 107296. https://doi.org/10.1016/j.chb.2022.107296

Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, et al. Bias and Fairness in Large Language Models: A Survey. Computational Linguistics. 2024; 50: 1097–1179. https://doi.org/10.1162/coli_a_00524

Ganguli D, Askell A, Schiefer N, Liao TI, Lukošiūtė K, Chen A, et al. The Capacity for Moral Self-Correction in Large Language Models. arXiv. 2023. https://doi.org/10.48550/arXiv.2302.07459 (preprint)

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv. 2024. https://doi.org/10.48550/arXiv.2312.10997 (preprint)

Gauthreaux K, Bonnett TA, Besser LM, Brenowitz WD, Teylan M, Mock C, et al. Concordance of Clinical Alzheimer Diagnosis and Neuropathological Features at Autopsy. Journal of Neuropathology and Experimental Neurology. 2020; 79: 465–473. https://doi.org/10.1093/jnen/nlaa014

General Medical Council. Good medical practice 2024. 2024. Available at: https://www.gmc-uk.org/professional-standards/good-medical-practice-2024 (Accessed: 6 July 2024).

Gove D, Nielsen TR, Smits C, Plejert C, Rauf MA, Parveen S, et al. The challenges of achieving timely diagnosis and culturally appropriate care of people with dementia from minority ethnic groups in Europe. International Journal of Geriatric Psychiatry. 2021; 36: 1823–1828. https://doi.org/10.1002/gps.5614

Guo Z, Ling Z, Li Y. Detecting Alzheimer's Disease from Continuous Speech Using Language Models. Journal of Alzheimer's Disease. 2019; 70: 1163–1174. https://doi.org/10.3233/JAD-190452

He Y, Zamani E, Yevseyeva I, Luo C. Artificial Intelligence-Based Ethical Hacking for Health Information Systems: Simulation Study. Journal of Medical Internet Research. 2023; 25: e41748. https://doi.org/10.2196/41748

Health Research Authority. GDPR guidance, Health Research Authority. 2018. Available at: https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/data-protection-and-information-governance/gdpr-guidance/ (Accessed: 6 July 2024).

Hossain G, Pomare Z S, Prybutok G. 'ChatGPT: A Companion for Dementia Care', 2024 IEEE International Conference on Consumer Electronics (ICCE) (pp. 1–6). Las Vegas, NV, USA. 28 February 2024. IEEE: New York, NY, USA. 2024. https://doi.org/10.1109/ICCE59016.2024.10444253

Hughey R, Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Bioinformatics. 1996; 12: 95–107. https://doi.org/10.1093/bioinformatics/12.2.95.

IBM. What Are AI Hallucinations? 2023. Available at: https://www.ibm.com/topics/ai-hallucinations (Accessed: 6 July 2024).

Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021; 37: 2112–2120. https://doi.org/10.1093/bioinformatics/btab083

Jiang Z, Araki J, Ding H, Neubig G. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. Transactions of the Association for Computational Linguistics. 2021; 9: 962–977. https://doi.org/10.1162/tacl_a_00407

Jiao B, Li R, Zhou H, Qing K, Liu H, Pan H, et al. Neural biomarker diagnosis and prediction to mild cognitive impairment and Alzheimer's disease using EEG technology. Alzheimer's Research & Therapy. 2023; 15: 32. https://doi.org/10.1186/s13195-023-01181-1

Knapp M, Shehaj X, Wong G, Hall A, Hanratty B, Robinson L, et al. Digital technology to support people living with dementia and carers. NIHR Older People and Frailty Policy Research Unit, p. 40. 2022. Available at: https://documents.manchester.ac.uk/display.aspx?DocID=60761 (Accessed: 6 July 2024).

Koychev I, Harrison J, Malhotra P, Dunne R, Sheehan B. Shifting paradigms in dementia care: navigating new therapies and prevention strategies. The British Journal of Psychiatry. 2024; 224: 187–188. https://doi.org/10.1192/bjp.2024.75

Lauscher A, Lueken T, Glavaš G. Sustainable modular debiasing of language models. In Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics. 2021; 4782–4797. https://doi.org/10.18653/v1/2021.findings-emnlp.411

Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. The Lancet Digital Health. 2023; 5: e333–e335. https://doi.org/10.1016/S2589-7500(23)00083-3

Li R, Wang X, Lawler K, Garg S, Bai Q, Alty J. Applications of artificial intelligence to aid early detection of dementia: A scoping review on current capabilities and future directions. Journal of Biomedical Informatics. 2022; 127: 104030. https://doi.org/10.1016/j.jbi.2022.104030

Lin S, Hilton J, Evans O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. 2022; 1: 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

Lin W, Gao Q, Yuan J, Chen Z, Feng C, Chen W, et al. Predicting Alzheimer's Disease Conversion From Mild Cognitive Impairment Using an Extreme Learning Machine-Based Grading Method With Multimodal Data. Frontiers in Aging Neuroscience. 2020; 12: 77. https://doi.org/10.3389/fnagi.2020.00077

Liu J, Yang M, Yu Y, Xu H, Li K, Zhou X. Large language models in bioinformatics: applications and perspectives. arXiv. 2024. (preprint)

Liu NF, Zhang T, Liang P. Evaluating Verifiability in Generative Search Engines. In Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics. 2023; 1: 7001–7025. https://doi.org/10.18653/v1/2023.findings-emnlp.467

Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. The Lancet. Digital Health. 2022; 4: e384–e397. https://doi.org/10.1016/S2589-7500(22)00003-6

Medicines and Healthcare Products Regulatory Agency. AI Airlock: the regulatory sandbox for AIaMD. 2024. Available at: https://www.gov.uk/government/collections/ai-airlock-the-regulatory-sandbox-for-aiamd (Accessed: 6 July 2024).

Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: A scoping review. iScience. 2024; 27: 109713. https://doi.org/10.1016/j.isci.2024.109713

Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digital Medicine. 2023; 6: 120. https://doi.org/10.1038/s41746-023-00873-0

Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. ACM Computing Surveys. 2023; 56: 30:1–30:40. https://doi.org/10.1145/3605943

Minssen T, Vayena E, Cohen IG. The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models. JAMA. 2023; 330: 315–316. https://doi.org/10.1001/jama.2023.9651

Misra R, Keane PA, Hogg HDJ. How should we train clinicians for artificial intelligence in healthcare? Future Healthcare Journal. 2024; 11: 100162. https://doi.org/10.1016/j.fhj.2024.100162

Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach. AI and Ethics. 2024; 4: 1085–1115. https://doi.org/10.1007/s43681-023-00289-2

NHS England. Medical devices and digital tools. 2023. Available at: https://www.england.nhs.uk/long-read/medical-devices-and-digital-tools/ (Accessed: 6 July 2024).

Obradovich N, Khalsa SS, Khan W, Suh J, Perlis RH, Ajilore O, et al. Opportunities and Risks of Large Language Models in Psychiatry. NPP - Digital Psychiatry and Neuroscience. 2024; 2: 8. https://doi.org/10.1038/s44277-024-00010-z

Ong JCL, Chang SYH, William W, Butte AJ, Shah NH, Chew LST, et al. Ethical and regulatory challenges of large language models in medicine. The Lancet. Digital Health. 2024; 6: e428–e432. https://doi.org/10.1016/S2589-7500(24)00061-X

Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, Ferri CP. The global prevalence of dementia: a systematic review and metaanalysis. Alzheimer's & Dementia. 2013; 9: 63–75.e2. https://doi.org/10.1016/j.jalz.2012.11.007

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. 2019. Available at: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2019573cc28650d14dfe (Accessed: 4 July 2024).

Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. Journal of the American Medical Informatics Association. 2020; 27: 491–497. https://doi.org/10.1093/jamia/ocz192

Reverberi C, Rigon T, Solari A, Hassan C, Cherubini P, GI Genius CADx Study Group, et al. Experimental evidence of effective human-AI collaboration in medical decision-making. Scientific Reports. 2022; 12: 14952. https://doi.org/10.1038/s41598-022-18751-2

Saeidnia HR, Kozak M, Lund BD, Hassanzadeh M. Evaluation of ChatGPT's responses to information needs and information seeking of dementia patients. Scientific Reports. 2024; 14: 10273. https://doi.org/10.1038/s41598-024-61068-5

Sezgin E. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. Digital Health. 2023; 9: 20552076231186520. https://doi.org/10.1177/20552076231186520

Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019; 25: 44–56. https://doi.org/10.1038/s41591-018-0300-7

Treder MS, Lee S, Tsvetanov KA. Introduction to Large Language Models (LLMs) for dementia care and research. Frontiers in Dementia. 2024; 3: 1385303. https://doi.org/10.3389/frdem.2024.1385303

Tripathi S, Gabriel K, Tripathi PK, Kim E. Large language models reshaping molecular biology and drug development. Chemical Biology & Drug Design. 2024; 103: e14568. https://doi.org/10.1111/cbdd.14568

Tsamakis K, Gadelrab R, Wilson M, Bonnici-Mallia AM, Hussain L, Perera G, et al. Dementia in People from Ethnic Minority Backgrounds: Disability, Functioning, and Pharmacotherapy at the Time of Diagnosis. Journal of the American Medical Directors Association. 2021; 22: 446–452. https://doi.org/10.1016/j.jamda.2020.06.026

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. Advances in Neural Information Processing Systems. 2017; 30: 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349

Wang W, Ferrari D, Haddon-Hill G, Curcin V. Electronic Health Records as Source of Research Data. In Colliot O (ed.) Machine Learning for Brain Disorders (pp. 331–354). Springer US (Neuromethods): New York, NY. 2023. https://doi.org/10.1007/978-1-0716-3195-9_11.

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems. 2022; 35: 24824–24837.

Xie B, Tao C, Li J, Hilsabeck RC, Aguirre A. Artificial Intelligence for Caregivers of Persons With Alzheimer's Disease and Related Dementias: Systematic Literature Review. JMIR Medical Informatics. 2020; 8: e18189. https://doi.org/10.2196/18189

Xue C, Kowshik SS, Lteif D, Puducheri S, Jasodanand VH, Zhou OT, et al. AI-based differential diagnosis of dementia etiologies on multimodal data. Nature Medicine. 2024; 30: 2977–2989. https://doi.org/10.1038/s41591-024-03118-z

Yang Y, Liu X, Jin Q, Huang F, Lu Z. Unmasking and quantifying racial bias of large language models in medical report generation. Communications Medicine. 2024; 4: 176. https://doi.org/10.1038/s43856-024-00601-z

Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. Advances in Neural Information Processing Systems. 2024; 36.

Yaros O, Hajda O, Prinsley MA, Randall R, Hepworth E. UK Government proposes a new approach to regulating artificial intelligence (AI). 2022. Available at: https://www.mayerbrown.com/en/insights/publications/2022/08/uk-government-proposes-a-new-approach-to-regulating-artificial-intelligence-ai (Accessed: 6 July 2024).

Yasunaga M, Chen X, Li Y, Pasupat P, Leskovec J, Liang P, et al. Large Language Models as Analogical Reasoners. arXiv. 2024. https://doi.org/10.48550/arXiv.2310.01714 (preprint)

Ye H, Liu T, Zhang A, Hua W, Jia W. Cognitive Mirage: A Review of Hallucinations in Large Language Models. The First International OpenKG Workshop: Large Knowledge-Enhanced Models. 2024. Available at: https://ceur-ws.org/Vol-3818/paper2.pdf (Accessed: 6 July 2024).

Yuan J, Tang R, Jiang X, Hu X. Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching. AMIA ... Annual Symposium Proceedings. AMIA Symposium. 2024; 2023: 1324–1333.

Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. The Lancet. Digital Health. 2024; 6: e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X

Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv. 2023. https://doi.org/10.48550/arXiv.2309.01219 (preprint)

Zheng Y, Gan W, Chen Z, Qi Z, Liang Q, Yu PS. Large Language Models for Medicine: A Survey. International Journal of Machine Learning and Cybernetics. 2025; 16: 1015–1040. https://doi.org/10.1007/s13042-024-02318-w