Received: 12 September 2024 | Revised: 28 September 2024 | Accepted: 13 October 2024

DOI: 10.1002/mpr.70007

LONG DIDACTIC PAPER

WILEY

From statistics to deep learning: Using large language models in psychiatric research

Yining Hua[1,2] | Andrew Beam[1,3] | Lori B. Chibnik[1,4] | John Torous[2,5]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[2]Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

[3]The CAUSALab, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[4]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

[5]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, USA

Correspondence
John Torous, Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Ave, Boston, MA, 02446, USA.
Email: jtorous@bidmc.harvard.edu

Abstract

Background: Large Language Models (LLMs) hold promise in enhancing psychiatric research efficiency. However, concerns related to bias, computational demands, data privacy, and the reliability of LLM-generated content pose challenges.

Gap: Existing studies primarily focus on the clinical applications of LLMs, with limited exploration of their potentials in broader psychiatric research.

Objective: This study adopts a narrative review format to assess the utility of LLMs in psychiatric research, beyond clinical settings, focusing on their effectiveness in literature review, study design, subject selection, statistical modeling, and academic writing.

Implication: This study provides a clearer understanding of how LLMs can be effectively integrated in the psychiatric research process, offering guidance on mitigating the associated risks and maximizing their potential benefits. While LLMs hold promise for advancing psychiatric research, careful oversight, rigorous validation, and adherence to ethical standards are crucial to mitigating risks such as bias, data privacy concerns, and reliability issues, thereby ensuring their effective and responsible use in improving psychiatric research.

KEYWORDS
artificial intelligence, clinical psychiatry, large language models, machine learning, psychiatric epidemiology, psychiatry

1 | INTRODUCTION

The integration of Large Language Models (LLMs) into psychiatry and mental health research is poised to revolutionize the field by enhancing diagnostic accuracy, personalizing care, and streamlining administrative tasks (Hua et al., 2024; Obradovich et al., 2024; Zhou et al., 2023). LLMs can efficiently process vast amounts of data, summarize clinical notes, and assist in complex decision-making processes (Yu et al., 2023). The rapid adoption of LLMs also brings unavoidable challenges. The unpredictability of LLM outputs, their potential to reinforce biases, and the risk of over-reliance on these models by clinicians are critical concerns. For example, while LLMs can suggest potential treatments, they may also propose suboptimal or contraindicated options, underscoring the need for careful human oversight. Additionally, the privacy implications of using LLMs in psychiatry are particularly concerning due to the sensitive nature of psychiatric records, the stigma and potential discrimination associated with mental health issues, and the critical importance of trust in therapeutic relationships, all of which heighten the risks associated with managing and potentially leaking sensitive patient data.

Existing reviews and perspective studies on LLMs in psychiatry have largely focused on their integration into clinical settings, missing the potential of LLMs in the broader context of psychiatric research: Obradovich et al. explored the ethical challenges and potential benefits of LLMs in psychiatric care, particularly in enhancing diagnostic accuracy and streamlining clinical processes (Obradovich et al., 2024); Volkmer et al. delved into the technical aspects of LLMs, emphasizing their architecture, potential clinical applications, and associated biases and privacy concerns (Volkmer et al., 2024); Omar et al. provided a systematic review of the practical applications of LLMs in psychiatry, highlighting their roles in clinical reasoning and diagnostic support particularly in complex and high-risk scenarios (Omar et al., 2024). The current study, in comparison, considers the uses of LLMs beyond clinical applications in the areas of literature review, study design, subject selection, statistical modeling, and academic writing. By addressing these areas, we seek to fill the gap in the current literature by demonstrating the broader research capabilities of LLMs in psychiatry, offering new insights into how these models can enhance the efficiency and accuracy of psychiatric research methodologies.

2 | BACKGROUND

2.1 | From statistics to deep learning

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that emerged from a diverse set of disciplines, including computer science, neuroscience, and formal logic. While statistical models are designed to summarize and explain relationships between variables in a dataset, ML models are created to enable computers to learn from data and improve their performance on specific tasks over time. This process is akin to how statistical models are fitted to data, but ML places a stronger emphasis on prediction and automated pattern recognition.

The core premise of ML is building systems that learn from experience and exposure to data. This objective naturally intersects with statistics, which focuses on collecting, analyzing, and drawing conclusions from data. However, unlike traditional statistical methods, which aim to generate human insights and explanations, ML often prioritizes the development of algorithms that can perform tasks such as classification, regression, and clustering without explicit instructions from humans.

In psychiatric epidemiology and services research, statistical models typically rely on predetermined equations where the form of the relationship between variables is specified based on theoretical knowledge or previous research. ML, on the other hand, uses algorithms that can automatically identify patterns in data without predefined equations. For example, optimization algorithms like stochastic gradient descent are used to train ML models by efficiently searching for optimal parameter values. This process is similar to the iterative fitting procedures in statistical models but allows for handling much larger and more complex datasets.

Deep learning (DL) is a subfield of ML that utilizes a model structure called neural networks, which are a layered architecture (hence “deep”) loosely inspired by the structure and function of the human brain. Neural networks consist of interconnected computational units called "neurons'' that apply weighted sums and nonlinear functions to transform input data hierarchically, with each neuron's output connected to subsequent layers. This process is analogous to the linear combination of predictors in a regression model, where each predictor is assigned a weight, and the sum of the weighted predictors is transformed by a link function (e.g., logit or probit) to obtain the outcome. However, in neural networks, these transformations occur in multiple layers, enabling the extraction of increasingly abstract features for tasks like classification and regression. The increased number and depth of parameters and thus increased capability in fitting data allows DL to excel in analyzing unstructured data such as images, audio, and texts with little human instruction or a priori distributional assumptions, pre-specified model equations and parameters, or explicit understanding of the resulting coefficients (Beam and Kohane, 2018). Meanwhile, the complex structure of DL models makes the training process more computationally intensive, often necessitating advanced computing resources such as graphics processing units (GPUs), tensor processing units (TPUs), or even large clusters of computing servers for fast parallel computation.

2.2 | Natural language processing and LLMs

Although LLMs now often exhibit the ability to process language, image, and even audio data, their origins lie in the evolution of natural language processing (NLP), a field specifically focusing on the interaction between computers and human language. NLP encompasses a wide range of techniques and approaches for analyzing, understanding, and generating human language. The field has evolved from early rule-based systems in the 1950s to the current ML-based approaches. Initially, the term “language model” referred to a specific type of mathematical function called a “loss function,” which measures how well the model's predictions match the actual data during the model training process. Nowadays, people use the term more broadly to refer to models capable of processing language.

Transformers, introduced in 2017, are a groundbreaking model architecture designed to handle large and complex data sets efficiently. Unlike earlier model architectures, Transformers use a self-attention mechanism that considers the entire input simultaneously rather than sequentially. This innovation allows Transformers to process information more effectively, making them particularly powerful for understanding and generating human language. Transformer-based models like OpenAI's GPT series, Stanford's LLaMa models, Google's PaLM, and Anthropic's Claude leverage this architecture to achieve high performance in various applications.

From a literal perspective, LLMs are “large,” but there is no unified standard defining how large a model must be to qualify as

an LLM. Researchers have found that scaling the model size and training data volume often enhances performance. This principle, known as the scaling law, is foundational to LLMs. Most LLMs are trained on as much data as possible. For example, GPT-4 was trained on corpora comprising roughly 10 trillion words. In what sometimes seems like magic, the large sizes of LLMs and extensive training data enable them to tackle complex tasks, such as writing poems and performing everyday tasks, simply by predicting the next words. However, training with more data and expanding model sizes are not a panacea. For instance, models can be prohibitively expensive to train and deploy, and they may replicate biases present in the training data. Much of the data used to train LLMs comes from social media, which often contains misinformation and false information. Recent research has focused on addressing various issues in LLMs, such as reducing model and training data size while maintaining performance and minimizing model hallucination (i.e., generating seemingly plausible but incorrect information).

It's not just about training data—researchers and developers also strive to align LLMs with human-like behavior rather than merely replicating Internet data. This process, known as Reinforcement Learning with Human Feedback (RLHF), involves researchers providing feedback during training to help models generate more accurate and contextually appropriate responses. Other research directions focus on enhancing LLM performance through various approaches, such as adding modules for retrieving reliable, relevant information instead of relying solely on the LLMs' memory.

The ability of LLMs to interact with humans via human-language interactions opens up new opportunities for psychiatrists and other mental health researchers who can now leverage the power of LLMs without necessarily having extensive expertise in ML or NLP. It blurs the line between research and engineering. Researchers can now "program" the model using human language instructions, rather than writing complex code or training the model from scratch. This democratizes access to powerful language models and allows domain experts to focus on their research questions rather than the intricacies of model development. However, this new paradigm also presents challenges. Users still need to understand how LLMs function and format their prompts in a way that the models can understand and process effectively. This requires a new set of skills, such as prompt engineering, which involves crafting prompts that elicit desired behaviors and outputs from LLMs.

3 | OPPORTUNITIES

3.1 | Literature review and study design

To design a robust psychiatric study, researchers must perform a literature review, variable selection, and data collection planning. Oftentimes, despite expert efforts to select suitable cohorts and control groups while minimizing biases, inherent human cognitive biases can influence research. For instance, psychiatric studies can be influenced by the "scientific period effect", where the current scientific understanding can limit or bias hypotheses and data collection choices, reflecting constraints of prevailing knowledge on research scope (Susser et al., 2000; Wadsworth et al., 2003). In addition, confirmation bias, a tendency to favor information that confirms existing beliefs, complicates objective decision-making, leading individuals to favor information that aligns with their existing beliefs.

In this context, LLMs emerge as powerful tools to mitigate such biases through their ability to suggest potential variables, relationships, and experimental designs in a conversational style (Dwivedi et al., 2023; Pournaras, 2023; van Dis et al., 2023). These models can widen exposure to a diverse range of literature and perspectives, by offering summarization of literature, possibly introducing disconfirming or alternative evidence, and prompting reconsideration of initial assumptions. Furthermore, LLMs can be explicitly guided to select viewpoints relevant to specific research needs (Antu et al., 2023). One approach is to use LLMs with Internet access, such as Bing's AI bot based on GPT models, or specialized academic search engines like Perplexity.ai, to identify and retrieve studies that meet specific criteria for a preliminary exploration. Given that most AI tools have limited search capabilities and their effectiveness depends on the underlying search engine (often Google or Bing), it remains prudent for researchers to conduct their own literature searches using established and comprehensive strategies.

Alternatively, LLMs without Internet access can be employed. In this method, researchers first compile a collection of loosely relevant studies and then use LLMs to screen and analyze this corpus. This technique enables the generation of customized literature reviews that align closely with the study's specific focus and requirements. For example, an empirical study (Guo et al., 2024a) demonstrates how GPT models can improve the efficiency and accuracy of screening titles and abstracts in clinical research reviews. By automating a traditionally manual process with OpenAI's GPT API, they found that GPT models achieved an average accuracy of 0.91, with a high agreement with human reviewers ($\kappa = 0.96$) on over 24,000 documents. They also proved that the capability of generative LLMs to justify their selections and adjust decisions enhances both the speed and quality of literature reviews. This highlights the potential of LLMs to support more precise and informed psychiatry research outcomes.

Nevertheless, the review of psychiatric papers is especially complex because clinical work with patients suffering from mental disorders is more interdisciplinary than in other areas of medicine. This complexity highlights the need for more guideline works to standardize and refine the review processes in psychiatric research, ensuring the integration of diverse disciplinary perspectives and reducing bias in this highly sensitive field.

3.2 | Finding study subjects and collecting study information

Subject selection in psychiatry is particularly challenging due to the stigma associated with mental health, which can lead to reluctance among potential participants to come forward, and difficulties in ensuring diverse and representative samples. This issue is compounded by the often complex and overlapping nature of psychiatric conditions, making it difficult to find subjects who meet the specific criteria for a study. LLMs have shown promise in automating the process of matching patients to trials by analyzing large datasets, including electronic health records and patient-reported outcomes, to identify suitable candidates. This can streamline recruitment, reduce biases in participant selection, and ensure a more precise match to the study's eligibility criteria, ultimately facilitating the recruitment of appropriate study subjects for psychiatric research. A relevant study highlighted the potential of Med-PaLM 2, an LLM trained on medical knowledge, in predicting psychiatric functioning from patient interviews and clinical descriptions (Galatzer-Levy et al., 2023). In a sample of 145 depression assessments and 115 PTSD assessments, Med-PaLM 2 achieved an accuracy range of 0.80–0.84 in predicting depression scores, demonstrating performance comparable to human clinical raters. This suggests that LLMs could aid in the assessment and selection of psychiatric study participants under proper guidance.

LLMs have also been used to translate complex medical information, such as diagnostic reports and study eligibility criteria, into plain language summaries. For instance, a recent study by Guo et al. demonstrated the potential of LLMs including GPT-4 and Llama-2 to simplify biomedical literature and generate lay language summaries (Guo et al., 2024b). Although the study was not specific to psychiatry, its findings highlight how translating medical jargon into accessible language can improve communication between researchers, healthcare providers, and potential participants. This enhanced communication is likely to increase understanding and participation willingness in psychiatric studies, as well as enable healthcare providers to more effectively identify and refer eligible patients.

Additionally, LLMs can help psychiatrists by turning unstructured patient data into clear, organized summaries. Psychiatric evaluations often involve complex patient histories that are difficult to simplify and analyze. LLMs can automate this process by converting free-text notes into structured summaries that highlight important diagnostic information and symptoms. This helps ensure that critical details are captured, leading to more accurate assessments and better-informed decisions. For example, a study involving North Korean defectors, who faced traumatic experiences, used LLMs including GPT-4 Turbo, GPT-3.5 Turbo, and Med-PaLM 2, to analyze their interview transcripts (So et al., 2024). These individuals often struggle with mental health issues like PTSD and depression. The LLMs were able to identify key symptoms such as anxiety and depression and summarize the stressors, including past trauma and current social challenges. This helped mental health practitioners better understand the patients' conditions and make more informed assessments. This study shows how LLMs can be practically applied in psychiatric settings to improve the accuracy and efficiency of mental health evaluations.

3.3 | Building statistical models and code generation

Building statistical models and writing code are crucial aspects of automating data processing and analysis, as well as effectively communicating findings. Unlike software engineering, coding in psychiatric studies often focuses on statistical modeling, which requires transforming complex mathematical formulations and study designs into efficient code.

Currently, there is no benchmark or dataset specifically designed to evaluate the performance of LLMs in generating statistical modeling code based on study designs or problem statements. However, a recent review study (Zheng et al., 2024) comparing LLMs in terms of their coding abilities found that for languages commonly used in statistical modeling and data science, such as Python and *R*, models fine-tuned on data science-specific datasets such as CodeT5 (Wang et al., 2021) and JuPyT5 (Chandel et al., 2022), fine-tuned variants of the LLM "Text-To-Text Transfer Transformer" (T5) (Raffel et al., 2020), offer the best performance. These models can generate code from descriptions, identify and correct errors, optimize performance, and provide explanations for code logic, making them invaluable tools for statistical modeling.

In addition to fine-tuned models, general-purpose models like GPT-3 and GPT-4 can also aid in data analysis by leveraging their extensive training in code and scientific literature. GitHub Copilot, powered by Codex, enhances coding efficiency by suggesting code snippets and documentation, supporting a wide range of programming tasks, including data science in Python. Another notable model is GPT-NeoX (Black et al., 2022), an open-source 20B-parameter LLM. It shows impressive performance on mathematical tasks though the original paper does not provide experimental evaluations of its programming capabilities. Already in 2024 there are efforts to use this model for mental health research, but results are still nascent (Sharp et al., 2023).

Although most of the attention in the field of code generation has been focused on software engineering, LLMs can still help psychiatric researchers with various aspects of coding, improving efficiency in tasks such as explaining code, translating code between languages, and automating code documentation (Meyer et al., 2023). They also offer automation beyond code generation, assisting with data cleaning, transformation, analysis, and visualization. By clarifying and suggesting enhancements to existing code, LLMs facilitate better understanding, maintenance, and improvement of research codebases, which could facilitate psychiatry research and care.

3.4 | Writing papers

LLMs are extensively utilized to enhance syntactic language, particularly aiding non-native speakers in writing, proofreading, and structuring fragmented drafts into coherent articles. This ability allows researchers to dedicate more attention to innovation and the essence of their research rather than the linguistic presentation of their findings. They can be used to streamline the preparation of research manuscripts, ensuring clear communication of complex data and findings. This supports researchers in conveying their insights more effectively, facilitating a better understanding of public health trends and interventions.

It's evident that the use of AI in academic writing is being shaped by evolving guidelines. Journals are increasingly acknowledging the role of AI tools, including LLMs, with specific policies to ensure ethical and transparent use. For instance, Nature series journals state that AI tools cannot be credited as authors but encourage documenting the use of such tools in the methods section or acknowledgments (Artificial Intelligence (AI) Nature Portfolio; Gaggioli, 2023; Tools such as ChatGPT threaten, 2023). The Journal of the American Medical Association (JAMA) Network journals have issued guidance on the responsible employment of AI tools, emphasizing the need for transparent reporting of their use in manuscript preparation and research submissions (Flanagin et al., 2023). Notably, The New England Journal of Medicine (NEJM), under its NEJM AI initiative, not only permits but encourages the use of LLMs in submissions (Koller et al., 2023). This progressive stance by NEJM AI illustrates a supportive approach towards leveraging AI to enhance research and academic writing, reflecting a broader acceptance of AI's role in academic research with an emphasis on responsible use. However, it is not every journal that follows this trend to address this problem. A recent study on the use of LLMs in radiology journals found that nearly half of the top 50 radiology journals (44.9%) did not provide an explicit policy on LLM use within the author submission guidelines. Among the journals with explicit policies, there was considerable variation in disclosure requirements and locations. Eleven journals (40.7%) required authors to disclose LLM usage in a new dedicated section, while eight journals (29.6%) asked authors to include this information within the methods section. This variability highlights the lack of standardized practices across radiology journals (Lee et al., 2024). Researchers are encouraged to stay informed about the latest guidelines from their target journals and to transparently report the use of AI tools, ensuring that their use adheres to the highest standards of research integrity and ethics.

Despite the help of LLMs in improving academic writing, it is important to note that using LLMs to generate text in academic writing without careful review and fact-checking should be discouraged. LLMs generate seemingly new text based on the author's input, but this comes with risks. As probabilistic models trained on existing texts, LLMs can generate content that appears original but may contain inaccuracies, biases, or even plagiarized content. Therefore, any text generated by LLMs should be carefully reviewed and fact-checked by the authors to ensure its accuracy, originality, and adherence to ethical standards. This concern has also been reflected in journal policies: Science does not allow AI-generated text or figures in their published papers and prohibits naming ChatGPT as an author (Thorp, 2023). Similarly, Nature does not accept LLM tools as authors and requires researchers to document their use in the methods and/or acknowledgments section (Tools such as ChatGPT threaten, 2023).

3.5 | Enhancing academic peer review

LLMs have also been applied to evaluating academic literature, offering feedback that aligns with the standards of selected fields and journals. These models leverage extensive training datasets encompassing a vast corpus of published materials, enabling them to approximate the average quality of literature across various fields and serve as proxy reviewers for academic manuscripts.

In such applications, LLMs are typically tasked with assuming the role of editors from specific journals, such as Nature. Researchers can instruct the model to critique and provide feedback on manuscripts as if they were undergoing the peer review process, offering insights into their originality, rigor, and potential impact.

A notable study utilized an automated pipeline with GPT-4 to generate scientific feedback on full PDFs of research papers (Liang et al., 2023). This study quantitatively compared the feedback from GPT-4 with that of human reviewers across 15 Nature journals and the International Conference on Learning Representations conference, finding an overlap in feedback points that was comparable to the overlap between two human reviewers. Furthermore, over half of the researchers in a subsequent user study found the feedback from GPT-4 to be helpful, often more so than that from some human reviewers. This suggests that while LLM-generated feedback is not without its limitations—such as a tendency to focus excessively on specific experimental suggestions—it can complement human expertise, particularly when expert feedback is unavailable.

Nonetheless, this application remains rebated as the quality of LLM output heavily depends on the training data's quality and diversity. Biased or limited data can skew or incomplete the model's feedback. Additionally, LLMs lack the nuanced understanding and specialized expertise of human reviewers, particularly in highly specialized or emerging fields with scarce established literature (Hosseini and Horbach, 2023). Ethical considerations also arise concerning the reliance on automated systems for traditionally human-handled tasks, including potentially reduced accountability and transparency in the review process.

3.6 | Practical implementation

The practical implementation of LLMs in psychiatric research remains underexplored. For non-experts in AI and LLMs, navigating the technical complexities of integrating these models into their work can be challenging. One way to bridge this knowledge gap is for psychiatric researchers to engage in interdisciplinary learning, including participation in NLP conference workshops, training

TABLE 1 Large language models implementation considerations across stages of psychiatric care.

Stage	Support offered	Risk of harm	Data needed
Prevention	Personalized psychoeducation	Low	Access to high quality resources
Relapse/Onset detection	Risk prediction	Medium	Access to clinical training data
Diagnosis	Data driven assessment	Medium	Access to clinical diagnosis data
Treatment optimization	Data driven medication selection	Medium	Access to medication selection data
Emergency support	Crisis counseling	High	Access to crisis communications and outcomes
Maintenance support	Routine therapy support	Medium	Access to therapy sessions and outcomes

sessions, and collaborations with AI specialists. A recent review from our team highlighted various models, evaluation measures, and metrics of LLMs within the broader context of mental health research (Hua et al., 2024). However, related efforts remain scarce, and there is a need for more meaningful research in the practical implementation of these technologies in psychiatry.

To further explore this opportunity, we have outlined the various stages of psychiatric care where LLMs can be effectively applied (Table 1), understanding that more outcomes data is necessary to determine actual use cases. These include personalized psychoeducation, risk prediction, data-driven diagnosis, and emergency support. Implementing LLMs in these contexts requires not only access to specialized data—such as clinical diagnosis or medication selection data—but also an understanding of the associated risk of harm, which varies across applications. To make these tools accessible to non-experts, no-code or low-code platforms, such as OpenAI's API and Hugging Face's Model Hub, provide streamlined methods for researchers to leverage pre-trained LLMs without needing extensive programming expertise.

Access to high-quality clinical data is critical to the effectiveness of LLMs. Different stages of psychiatric care, from prevention to treatment optimization, require specialized data sources. Researchers may need to collaborate with clinical institutions or data providers to acquire the necessary data for fine-tuning models to their specific research questions. For instance, crisis counseling applications necessitate access to crisis communication data to mitigate the higher risk of harm associated with emergency interventions.

High-risk applications, such as emergency support, carry a greater risk of harm, making it essential to integrate human oversight into these systems. Researchers must adopt ethical guidelines to ensure that LLMs supplement, rather than replace, human judgment, particularly in sensitive, high-stakes scenarios like crisis support such as happened in 2023 with the TESSA chatbot designed to support users seeking help for eating disorders (Jiang et al., 2023).

4 | CHALLENGES

4.1 | Bias

Bias is a crucial challenge in LLMs due to the nature of their training data which generally lacks systematic selection and debiasing, predominantly coming from the Internet. This data can reflect societal biases and does not uniformly represent all demographic groups. Such biases can lead to LLMs disproportionately representing more frequently occurring groups, typically skewing towards white demographics. This bias is present not only in image recognition tasks but also in textual analysis, reflecting the broader biases inherent in Internet data. The lack of diverse contributions to online content further exacerbates this issue, potentially causing LLMs to misrepresent or misunderstand underrepresented groups. Continuous efforts to understand and mitigate these biases are crucial to developing fair and equitable AI systems.

4.2 | Computational limitations

Deploying LLMs requires substantial computational resources, often making direct training financially and technically impractical for many research groups. Given the extensive computational demands and associated costs, researchers often opt for pre-trained LLMs hosted on third-party platforms or APIs. This approach allows for the application of these models in zero-shot or few-shot learning scenarios, leveraging their broad inferential capabilities without the need for extensive retraining. Consequently, this strategy helps circumvent the high costs associated with custom model development.

While zero-shot or few-shot applications of LLMs have shown good performance in many medical fields, their effectiveness in psychiatric contexts remains underexplored and necessitates further empirical evaluation. This research gap underscores the need for future studies to assess the performance of LLMs in a range of subject matter areas. Such assessments are crucial to determine the feasibility of using LLMs more extensively in this field without incurring prohibitive costs.

4.3 | Data privacy

Data privacy is another core concern, especially given the computational limitations discussed in Section 4.2. Most mental health researchers rely on third-party LLM services rather than deploying models on local machines, which introduces the risk of data leakage. This issue is particularly critical when handling sensitive clinical data.

The European Union's General Data Protection Regulation mandates stringent measures to protect personal data and individual rights, setting a high standard for data privacy. Therefore, LLMs utilized in psychiatric research must navigate these regulations to prevent breaches of confidentiality.

The primary challenge lies in accessing the extensive datasets required for effective AI operation while ensuring compliance with data privacy laws and ethical standards. Implementing robust anonymization techniques and adhering to legal frameworks are essential, though complex. Additionally, there is a risk that LLMs might inadvertently reveal sensitive information, even from anonymized datasets, necessitating ongoing vigilance and innovation in data protection methods. In the context of psychiatry, where data include highly sensitive personal health information, the stakes are particularly high. A breach of data privacy can have severe consequences, not only for individuals but also for the credibility and ethical standing of the research community. Therefore, researchers must prioritize data privacy at every stage of LLM deployment, from data collection and preprocessing to model training and application.

Moreover, the ethical implications of using LLMs in psychiatry must be carefully considered. Researchers need to ensure that their use of these models aligns with the principles of beneficence, nonmaleficence, and justice. This involves being transparent about data usage, obtaining informed consent from data subjects whenever possible, and ensuring that the research benefits are distributed fairly.

4.4 | Validity and reliability

LLMs may exhibit shortcomings in delivering reliable responses, particularly in intricate and domain-specific tasks, with reliability concerns often stemming from the hallucination problem—wherein the model generates plausible yet fabricated information—an issue of notable concern within the literature. While extensive efforts have been dedicated to mitigating these issues, complete eradication remains elusive. Moreover, the generation process of LLMs typically involves the utilization of hyperparameters, such as temperature, to modulate the robustness of the generated context; however, when served by third parties, these hyperparameters are often configured to elicit varied responses, potentially resulting in divergent or conflicting answers to identical queries. Hence, researchers are advised to exercise vigilance in assessing the reliability of content generated by LLMs, advocating for thorough verification of the underlying sources of the generated contexts.

Validity poses another challenge, largely attributed to the inadequacy of relevant training data. For instance, LLMs commonly exhibit limitations in directly addressing complex inquiries, such as those encountered in biostatistics. In a study assessing the utility of ChatGPT (GPT-3.5 and GPT-4) as a study aid for solving biostatistical problems sourced from the Handbook of Medical Statistics by Peacock and Peacock, both versions achieved a 50 to 60 percent accuracy rate on initial attempts (Ignjatović and Stevanović, 2023). Furthermore, a recent benchmark investigation across various biomedical NLP tasks underscores the suboptimal performance of LLMs compared to more specialized, smaller-scale classification models tailored to specific tasks. These findings underscore the inherent limitations of LLMs in scenarios necessitating heightened diagnostic precision, emphasizing the importance of selecting appropriate models aligned with the specific requirements of the study at hand.

5 | Conclusion

The integration of LLMs into psychiatric and more general mental health disorders research offers advancements in efficiency and accuracy, particularly in literature review, study design, subject selection, and academic writing. However, their use also presents challenges, including potential biases, substantial computational demands, data privacy concerns, and issues with reliability and validity. To fully harness LLMs' potential while mitigating these risks, researchers must exercise careful oversight, ensure rigorous validation, and adhere to strict ethical standards, particularly in handling sensitive psychiatric data.

AUTHOR CONTRIBUTIONS

Yining Hua: Conceptualization; writing—original draft; writing—review & editing; methodology. **Andrew Beam**: Conceptualization; supervision; writing—review & editing; validation. **Lori B. Chibnik**: Supervision; conceptualization; writing—original draft; writing—review & editing; validation. **John Torous**: Supervision; validation; writing—review & editing.

CONFLICT OF INTEREST STATEMENT

JT has research support from Otsuka and is an adviser to Precision Mental Wellness. All other authors have no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Yining Hua https://orcid.org/0000-0001-7779-1208

REFERENCES

Antu, S. A., Chen, H., & Richards, C. K. (2023). Using LLM (Large Language model) to improve efficiency in literature review for undergraduate research. In S. Moore, J. C. Stamper, R. Tong, C. Cao, Z. Liu, X. Hu, Y. Lu, J. Liang, H. Khosravi, P. Denny, A. Singh, & C. Brooks. (Eds.), *Proc workshop empower educ LLMs - -gen interface content gener 2023 Co-located 24th int conf artif intell educ AIED 2023 Tokyo jpn july 7 2023* (pp. 8–16). CEUR-WS.org.

Artificial Intelligence (AI) Nature Portfolio. [Internet]. Retrieved Feb 26, 2024 from https://www.nature.com/nature-portfolio/editorial-policies/ai

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, *319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., & Weinbach, S. (2022). GPT-NeoX-20B: An open-source autoregressive language model. In A. Fan, S. Ilic, T. Wolf, & M. Gallé (Eds.), *Proc BigScience episode 5 – workshop chall perspect creat large lang models [internet]* (pp. 95–136). virtual+Dublin: Association for Computational Linguistics. https://aclanthology.org/2022.bigscience-1.9

Chandel, S., Clement, C. B., Serrato, G., & Sundaresan, N. (2022). Training and evaluating a jupyter notebook data science assistant [internet]. *arXiv*. Retrieved Apr 4, 2024 from http://arxiv.org/abs/2201.12901

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., …, & Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Flanagin, A., Kendall-Taylor, J., & Bibbins-Domingo, K. (2023). Guidance for authors, peer reviewers, and editors on use of AI, Language Models, and chatbots. *JAMA*, *330*(8), 702–703. https://doi.org/10.1001/jama.2023.12500

Gaggioli, A. (2023). Ethics: Disclose use of AI in scientific manuscripts. *Nature*, *614*(7948), 413. https://doi.org/10.1038/d41586-023-00381-x

Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A., & Malgaroli, M. (2023). The capability of large Language Models to measure psychiatric functioning [internet]. *arXiv*. Retrieved Aug 5, 2024 from http://arxiv.org/abs/2308.01834

Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., & Naugler, C. (2024a). Automated paper screening for clinical reviews using large Language Models: Data analysis study. *Journal of Medical Internet Research*, *26*, e48996. https://doi.org/10.2196/48996

Guo, Y., Qiu, W., Leroy, G., Wang, S., & Cohen, T. (2024b). Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, *149*, 104580. https://doi.org/10.1016/j.jbi.2023.104580

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res Integr Peer Rev*, *8*(1), 4. https://doi.org/10.1186/s41073-023-00133-5

Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y., Zhou, P., Moran, L. V., Ananiadou, S., Beam, A., & Torous J. (2024). Large Language models in mental health care: A scoping review. *arXiv*.

Ignjatović, A., & Stevanović, L. (2023). Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: A descriptive study. *J Educ Eval Health Prof*, *20*, 28. https://doi.org/10.3352/jeehp.2023.20.28

Jiang, G., Xu, M., Zhu, S. C., Han, W., Zhang, C., & Zhu, Y. (2023). Evaluating and inducing personality in pre-trained language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Adv neural inf process syst* (Vol. 36, pp. 10622–10643). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/21f7b745f73ce0d1f9bcea7f40b1388e-Paper-Conference.pdf

Koller, D., Beam, A., Manrai, A., Ashley, E., Liu, X., Gichoya, J., Holmes, C., Zou, J., Dagan, N., Wong, T. Y., Blumenthal, D., & Kohane, I. (2023). Why we support and encourage the use of large Language Models in NEJM AI submissions. *NEJM AI*, *1*, AIe2300128. https://doi.org/10.1056/aie2300128

Lee, T.-L., Ding, J., Trivedi, H. M., Gichoya, J. W., Moon, J. T., & Li, H. (2024). Understanding radiological journal views and policies on large Language Models in academic writing. *Journal of the American College of Radiology*, *21*(4), 678–682. https://doi.org/10.1016/j.jacr.2023.08.001

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., & Zou, J. (2023). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. Retrieved May 28, 2024 from https://arxiv.org/abs/2310.01783

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Mining*, *16*(1), 20. https://doi.org/10.1186/s13040-023-00339-9

Obradovich, N., Khalsa, S. S., Khan, W. U., Suh, J., Perlis, R. H., Ajilore, O., & Paulus, M. P. (2024). Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry Neurosci*, *2*, 1–8. https://doi.org/10.1038/s44277-024-00010-z

Omar, M., Soffer, S., Charney, A. W., Landi, I., Nadkarni, G. N., & Klang, E. (2024). Applications of large language models in psychiatry: A systematic review [internet]. *Frontiers in Psychiatry*, *15*. https://doi.org/10.3389/fpsyt.2024.1422807

Pournaras, E. (2023). Science in the era of ChatGPT, large language models and generative AI: Challenges for research ethics and how to respond (pp. 275–290).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 5485–5551.

Sharp, G., Torous, J., & West, M. (2023). Ethical challenges in AI approaches to eating disorders. *Journal of Medical Internet Research*, *25*, e50696. https://doi.org/10.2196/50696

So, J., Chang, J., Kim, E., Na, J., Choi, J., Sohn, J., Kim, B., & Chu, S. H. (2024). Aligning large Language Models for enhancing psychiatric interviews through symptom delineation and summarization [internet]. *arXiv*. Retrieved Aug 5, 2024 from http://arxiv.org/abs/2403.17428

Susser, E., Terry, M. B., & Matte, T. (2000). The birth cohorts grow up: New opportunities for epidemiology. *Paediatric & Perinatal Epidemiology*, *14*(2), 98–100. https://doi.org/10.1046/j.1365-3016.2000.00249.x

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science.*, *379*(6630), 313. https://doi.org/10.1126/science.adg7879

Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*. 2023;*613*:612.

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

Volkmer, S., Meyer-Lindenberg, A., & Schwarz, E. (2024). Large language models in psychiatry: Opportunities and challenges. *Psychiatry Research*, *339*, 116026. https://doi.org/10.1016/j.psychres.2024.116026

Wadsworth, M. E. J., Butterworth, S. L., Hardy, R. J., Kuh, D. J., Richards, M., Langenberg, C., Hilder, W., & Connor, M. (2003). The life course prospective design: An example of benefits and problems associated with study longevity. *Social Science & Medicine*, *57*(11), 2193–2205. https://doi.org/10.1016/s0277-9536(03)00083-2

Wang, Y., Wang, W., Joty, S., & Hoi, S. C. H. (2021). CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proc 2021 conf empir methods nat lang process. Online and punta cana, Dominican Republic* (pp. 8696–8708). Association for Computational Linguistics.

Yu, P., Xu, H., Hu, X., & Deng, C. (2023). Leveraging generative ai and large Language Models: A comprehensive roadmap for healthcare integration. *Healthcare*, *11*(20), 2776. https://doi.org/10.3390/healthcare11202776

Zheng, Z., Ning, K., Wang, Y., Zhang, J., Zheng, D., Ye, M., & Chen, J. (2024). A survey of large Language Models for code: Evolution, benchmarking, and future trends [internet]. *arXiv*. Retrieved Apr 4, 2024 from http://arxiv.org/abs/2311.10372

Zhou, H., Liu, F., Gu, B., Zou, X., Huang, J., Wu, J., Li, Y., Chen, S. S., Zhou, P., Liu, J., Hua, Y., Mao, C., You, C., Wu, X., Zheng, Y., Clifton, L., Li, Z., Luo, J., & Clifton, D. A. (2023). A survey of large Language Models in medicine: Principles, applications, and challenges [internet]. *arXiv*. Retrieved Feb 2, 2024 from http://arxiv.org/abs/2311.05112

How to cite this article: Hua, Y., Beam, A., Chibnik, L. B., & Torous, J. (2025). From statistics to deep learning: Using large language models in psychiatric research. *International Journal of Methods in Psychiatric Research*, e70007. https://doi.org/10.1002/mpr.70007