Fluency without understanding: risks of large language models in mental healthcare

Bahaa Hassan,[1] Abedelrahman Ahmed[2]

BJPsych Bulletin (2026) Page 1 of 4, doi:10.1192/bjb.2026.10227

[1]Norfolk and Suffolk NHS Foundation Trust, Norwich, UK; [2]Norfolk and Suffolk NHS Foundation Trust, Norwich, UK

Correspondence to Bahaa Hassan (dr.bahaa.g@gmail.com)

Against the Stream articles tackle controversial issues. The idea is to challenge conventional wisdom and stimulate discussion. *BJPsych Bulletin* is not responsible for statements made by contributors and material in *BJPsych Bulletin* does not necessarily reflect the views of the Editor-in-Chief or the College.

First received 30 Jan 2026, final revision 26 Feb 2026, accepted 2 Mar 2026

© The Author(s), 2026. Published by Cambridge University Press on behalf of Royal College of Psychiatrists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Summary We always treat fluent language as a marker of intelligence and trustworthiness, often independent of factual accuracy. Large language models (LLMs) exploit this bias by producing confident, human-like texts that are perceived as intelligent and trustworthy, even when they lack accurate contextual understanding or are factually incorrect. This creates particular risks in mental healthcare, where communication, trust and context are central, and where errors are difficult to detect but highly consequential. This article examines how linguistic fluency shapes judgement, how LLMs amplify these effects and why their use in mental healthcare poses ethical and clinical dangers. It argues for strict limits on deployment, restricting LLMs to supervised, assistive tasks rather than clinical judgement.

Keywords Medical technology; machine learning methods; service development; health economics; ethics.

Humans have long perceived language as an indicator of intelligence. Throughout history, articulate language has influenced how people judge the quality and trustworthiness of ideas, sometimes more than the underlying facts. Political and religious leaders have long used carefully crafted rhetoric to build trust and loyalty, reducing the chance of substantive scrutiny.[1] In some religious traditions, the sacred text itself is taken as a miracle, partly because of its literary perfection, making it hard for believers to imagine that anyone could have written it but God.

Insights from psychology and linguistics demonstrate that the form of language strongly affects how people evaluate information. Subtle linguistic techniques such as presupposition can introduce ideas as if they were already accepted facts, reducing critical scrutiny of the message. The language of the message also matters; bilingual audiences can find the same news more or less believable depending on which language it is presented in, even when they are equally proficient in both.[2] In other words, fluent, well-packaged language can make weak or false content feel convincing and trustworthy.

Language similarly shapes judgement in healthcare. Patients tend to place greater trust in clinicians who speak with clear, standard-sounding language than in equally competent clinicians with foreign accents or non-standard speech patterns. Studies show that doctors with foreign accents are often perceived as less competent and less trustworthy, even when they provide identical information. In online medical consultations, the use of standard accents is associated with higher patient satisfaction and greater willingness to follow advice, largely through increased perceptions of competence.[3] These findings support the idea that language biases our sense of who is perceived as 'professional' or 'intelligent', often independently of actual expertise.

How LLMs amplify existing errors of judgement

Large language models (LLMs) such as ChatGPT and others amplify this ancient bias in a new way. Technically, LLMs are trained on massive data-sets to perform next-token prediction. They calculate the statistical probability of the next piece of text in a sequence and generate language one token at a time to produce text that is coherent and contextually appropriate. Through this process, they excel at high-fidelity imitation of human language, reproducing its

https://doi.org/10.1192/bjb.2026.10227 Published online by Cambridge University Press

outward form without engaging in the cognitive and contextual processes by which humans develop their understanding.

This technical fluency shapes how users perceive the model. LLMs produce confident, human-like language, respond plausibly to follow-up questions and mirror users' assumptions and emotional tone, encouraging people to perceive them as intentional, intelligent agents. Experimental work shows that when the same chatbot is presented using more subjective, human-like language rather than neutral, machine-like phrasing, users rate it as more trustworthy and higher in quality, even though the underlying system, information and accuracy are unchanged.[4] This effect reflects how linguistic cues such as emotion, self-reference and conversational warmth inflate perceived competence and credibility without any corresponding increase in accuracy.

A study of medical LLMs has shown that these systems can appear highly competent while failing in clinically meaningful ways. Models often perform well on benchmarks (standardised, test-style evaluations with limited context) but make serious reasoning errors or confidently incorrect recommendations in more realistic, expert-level clinical tasks, leading non-expert evaluators to rate such outputs as high quality, whereas domain experts identify substantial inaccuracies and safety risks.[5] Similarly, a recent evaluation of a consumer LLM triage tool found that performance worsened at the extremes of acuity, with more than half of emergency cases under-triaged and high failure rates in both non-urgent and emergency scenarios. Crisis responses to suicidal ideation were inconsistently triggered.[6]

Why mental healthcare is at particular risk

In mental healthcare, where communication is central and language is the main tool, there is particular cause for concern. Therapeutic work depends on building a safe, trusting relationship in which patients can disclose sensitive experiences and feel understood. Psychiatry and psychotherapy literature consistently show that seemingly small elements such as introductions, active listening, appropriate questioning, sensitivity to culture, and body language have a substantial impact on engagement and clinical outcomes. Improvements in clinicians' communication skills are associated with better patient experiences of the therapeutic relationship, whereas negative communication styles, including criticism, hostility and over-involvement, are linked to higher relapse rates in conditions such as schizophrenia. Good language alone is not sufficient, but poor or careless communication can undermine even clinically sound treatment.

Because therapeutic communication is both powerful and easily compromised, introducing LLMs into mental healthcare is especially risky. Patients seeking help are often vulnerable and suggestible, particularly when distressed, ashamed or isolated. LLMs are designed to be agreeable and reassuring; they frequently mirror users' assumptions and present outputs with unwarranted certainty, a behaviour described as sycophancy.[7] Although studies suggest potential benefits in areas such as screening or e-health support, they also emphasise that current risks in clinical use, including hallucinations, inconsistency, failure to recognise errors, lack of accountability, data privacy concerns and over-reliance, may outweigh these benefits.[8] This has led many authors to caution against using LLMs as substitutes for professional mental healthcare or deploying them unsupervised for clinical advice.

A further problem is that errors are usually identifiable only by those with relevant domain expertise, and are often effectively invisible to non-experts, allowing outputs to sound highly persuasive outside areas of the user's specialist knowledge. In domains such as programming or mathematics, tasks are more objective and errors typically lead to immediately costly consequences, which makes those errors easier to identify and correct. In medicine and mental healthcare, however, the 'ground truth' and the consequences of dysfunctional outputs or serious errors are often not immediately visible, making subtle inaccuracies harder to detect and their consequences far more serious, particularly over the longer term and when deployed at scale.

Beyond direct clinical risks, the widespread use of LLMs by professionals threatens to reinforce and amplify subtle errors and biases, as repeated exposure to confident but flawed outputs normalises mistakes and embeds them into routine practice. Over-reliance on these systems may also lead to the deskilling of the workforce, as professionals gradually lose the capacity for critical reading, precise expression and independent thinking. Research on LLMs in healthcare suggests that over-reliance may shift clinicians from active cognitive engagement to supervisory checking, reducing opportunities to practise core skills such as formulation, clinical reasoning and precise documentation. Over time, this creates a feedback loop in which diminished human expertise increases dependence on automated outputs, further eroding professional judgement and autonomy.[9]

Why efficiency-driven models fail mental healthcare

Assessment and treatment in mental healthcare rely on a biopsychosocial formulation that integrates psychological factors, socioeconomic context and biological factors, an approach to care that fits poorly with efficiency-driven models and magnifies the risks described above, as this complexity is reduced to checklists, metrics and simplified inputs and outputs. The example of UK general practice will be familiar to many readers. A system that prioritises speed, efficiency and cost reduction pressurises general practitioners to manage complex mental health problems in 10-min appointments, leaving little opportunity to understand the psychological, social, cultural and developmental context of a person's presentation, and increasing the risk of misdiagnosis and harmful management.[10] There is also a long history of harm from unqualified or poorly supervised therapists who rely on persuasive language while offering misguided or harmful treatment.

LLMs risk reproducing this same efficiency-driven pattern on an automated, industrial scale. They are

https://doi.org/10.1192/bjb.2026.10227 Published online by Cambridge University Press

promoted as tools to optimise speed and efficiency at a lower cost, but because they operate on decontextualised text, they are likely to miss essential context. They also cannot reliably track long-term therapeutic relationships, subtle shifts in risk, or the non-verbal elements of communication that often signal crisis. They can deliver apparently empathetic, insightful language without training, accountability or reliable error recognition. When patients are already vulnerable and suggestible, this combination creates a serious ethical hazard.

These risks are not merely theoretical. Recent findings indicate that current LLMs frequently misinterpret clinical scenarios and generate unsafe therapeutic responses. In structured tests, models expressed stigmatising views towards people with mental illness and at times reinforced delusional beliefs rather than challenging them, reflecting sycophantic tendencies and optimisation for agreeable responses at the expense of sound clinical reasoning. In February 2024, a 14-year-old boy died by suicide after prolonged interaction with a chatbot that allegedly failed to provide consistent crisis responses when he expressed suicidal thoughts. Such failures point to structural limitations in how LLMs operate, and even newer models do not reliably adhere to established therapeutic principles or manage high-risk situations appropriately.[11]

Where LLMs may help, and where they should not

None of this means that LLMs have no role in mental healthcare. They can perform reasonably well in language-based tasks such as transcription, translation, summarisation and information extraction, and many clinicians already use them for drafting, editing and literature review. In services overwhelmed by documentation and fragmented records, carefully governed use of LLMs could free clinician time for direct patient care. In the UK, General Medical Council guidance emphasises this limited, assistive use under human oversight, with transparency and clear accountability, rather than delegation of clinical judgement.[12] However, although LLMs are getting better with scaling and more training, hallucinations and bias remain unresolved problems. Responsible integration therefore requires strict governance, clear task boundaries, human verification of outputs, strong privacy protections and explicit accountability when errors occur.

Although future artificial intelligence (AI) systems may contribute more meaningfully to mental healthcare, current technologies are not yet safe or reliable enough to justify overambitious deployment. This push is occurring in a policy context that increasingly frames healthcare in terms of economic productivity. In the UK, the current Secretary of State for Health and Social Care, Wes Streeting, has announced that the Department of Health and Social Care will expand its focus to boost economic growth. In this context, public funds are already flowing to private AI firms on the basis of efficiency claims with weak supporting evidence. If this goes unchecked, we risk shaping our mental health systems around the priorities of industrialised care, with technology firms offering scalable products and managers prioritising numerical efficiency metrics over patients' real-life outcomes, under pressure to reduce costs. In this context, persuasive but unreliable tools risk being adopted with insufficient caution, where the consequences of error are highest.

About the authors

Bahaa Hassan is a Year 3 Core Trainee (CT3) in psychiatry with Norfolk and Suffolk NHS Foundation Trust, Norwich, UK. **Abed Abedelrahman** is a Year 3 Core Trainee (CT3) in psychiatry with Norfolk and Suffolk NHS Foundation Trust, Norwich, UK.

Data availability

Data availability is not applicable to this article as no new data were created or analysed in this article.

Author contributions

B.H. developed the concept and main argument, identified the literature and wrote the initial draft. A.A. contributed additional literature, refined the manuscript and assisted with revisions in response to peer review before agreement for submission.

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Declaration of interest

None.

References

1 Raman SK. *'Dr. Fox Effect' on Student Engagement, Cognitive Load and Learning Outcomes*. ResearchGate, 2026 (https://www.researchgate.net/publication/399284578_'Dr_Fox_Effect'_on_Student_Engagement_Cognitive_Load_and_Learning_Outcomes).

2 Erlich A, Aslett K, Graham S, Tucker JA. How language shapes belief in misinformation: a study among multilinguals in Ukraine. *J Exp Political Sci* 2026; **13**: 76–90.

3 Xie P, Long X, Mu W, Wang J, Wu Y. Standard accent vs. local accent: exploring the role of doctors' accent characteristics in online medical services. *Health Commun* 2025; **40**: 1363–80.

4 Pan W, Liu D, Meng J, Liu H. Human–AI communication in initial encounters: how AI agency affects trust, liking, and chat quality evaluation. *New Media Soc* 2025; **27**: 5822–47.

5 Alaa A, Hartvigsen T, Golchini N, Dutta S, Dean F, Raji ID, et al. Medical large language model benchmarks should prioritize construct validity. *arXiv [csCL]* [Preprint] 2023. Available from: https://arxiv.org/abs/2503.10694.

6 Ramaswamy A, Tyagi A, Hugo H, Jiang J, Jayaraman P, Jangda M, et al. ChatGPT Health performance in a structured test of triage recommendations. *Nat Med* [Epub ahead of print] 23 Feb 2026. Available from: https://doi.org/10.1038/s41591-026-04297-7.

7 Carro MV. Flattering to deceive: the impact of sycophantic behavior on user trust in large language model. *arXiv [csAI]* [Preprint] 2024. Available from: https://arxiv.org/abs/2412.02802.

8 Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. *JMIR Ment Health* 2024; **11**: e57400.

9 Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. *J Med Internet Res* 2024; **26**: e56764.

https://doi.org/10.1192/bjb.2026.10227 Published online by Cambridge University Press

10 Dwivedi K, Huddy V, Oliver P, Burton C. Complex mental health difficulties in primary care: a scoping review with thematic synthesis. *BJGP Open* 2025; 9: BJGPO.2024.0223.

11 Moore J, Grabb D, Agnew W, Klyman K, Chancellor S, Ong DC, et al. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *arXiv* [Preprint] 2025. Available from: https://ar5iv.labs.arxiv.org/html/2504.18412.

12 General Medical Council. *GMC: Artificial Intelligence and Innovative Technologies*. GMC, 2026 (https://www.gmc-uk.org/professional-standards/learning-materials/artificial-intelligence-and-innovative-technologies).

https://doi.org/10.1192/bjb.2026.10227 Published online by Cambridge University Press