Using Generative AI for Mental Health Documentation

Why Public Tools Fail Professional Standards and What Safe Use Requires

ConfideAI Research

December 2025

confideai.ai

Using Generative AI for Mental Health Documentation: Why Public Tools Fail Professional Standards and What Safe Use Requires

Abstract

The rapid adoption of large language models (LLMs) in clinical practice has created urgent pressures for mental health professionals to use publicly available, generic AI tools for documentation, assessment, and therapeutic support [1][2]. However, emerging empirical evidence, regulatory guidance, and ethical analysis indicate that such use—whether with identifiable or superficially de-identified patient data—commonly conflicts with established professional standards governing confidentiality, record-keeping, clinical responsibility, and patient safety [3][4][5]. This paper synthesises current regulatory guidance from professional bodies (e.g., AHPRA, national AI clinical guides), legal and privacy analysis, and peer-reviewed literature to argue that generic, publicly hosted LLMs are unsuitable for routine mental health documentation without robust health-sector-grade governance and technical safeguards [6][7]. The paper further highlights a critical barrier for independent practitioners: the absence of affordable, compliant alternatives leaves solo clinicians and small practices facing a de facto choice between unsafe tools and no tools, a situation that creates both professional indemnity risks and a widening gap in access to responsible AI-assisted practice. Practical guidance and "red lines" are proposed, along with principles for clinically governed AI environments that embed professional standards in technical design and organisational accountability.

1. Introduction

Large language models have achieved remarkable capabilities in text generation, summarisation, and question-answering, and their potential to streamline clinical workflows is evident [1][8]. Mental health professionals face substantial administrative burdens in documentation, treatment planning, and communication; the promise of AI-assisted scribing, note generation, and formulation support is therefore attractive [9]. Early commentaries and case reports describe efficiency gains and improved consistency when LLMs assist with routine documentation tasks [1][3].

However, systematic reviews of ChatGPT and other LLMs in healthcare settings, along with regulatory and ethical analyses published in 2023–2025, highlight serious unresolved concerns about privacy, reliability, bias, accountability, and the fit between generic LLM capabilities and the exacting standards of mental health practice [1][2][10][11][12]. In mental health, the stakes are particularly high. Clinical records are repositories of intimate information—trauma histories, abuse, relationship detail, sexual orientation, substance use, forensic risk—that patients entrust clinicians to protect absolutely [13][14].

Breaches or re-identification of such information can cause profound harm: stigma, discrimination, loss of employment, custody loss, safety risk, or re-traumatisation [13][5].

At the same time, independent practitioners and small mental health practices operate without IT departments, legal teams, or formal data-governance structures. They face a stark reality: if they wish to use any AI tool, they often have only two options: paste notes into a public chatbot (accepting opaque data handling and broad liability disclaimers), or forego AI entirely. Neither option is sustainable or professionally safe [15][16]. Commercial enterprise solutions exist, but they are often priced for large health systems, not solo clinicians [6][16].

This paper examines why generic LLMs, when deployed outside formally governed health-sector environments, fail to meet the professional and legal standards expected of mental health professionals. It then outlines what responsible AI use requires: technical safeguards, organisational governance, and a practical framework that independent practitioners can actually access and implement. The paper is aimed at mental health professionals (psychologists, psychiatrists, clinical counselors, social workers, and therapists) who wish to understand the risks, the standards, and the path toward defensible AI-assisted practice.

2. Professional Standards and Regulatory Expectations

2.1 Clinician Responsibility and Professional Obligations

Professional regulators and licensing bodies in Australia, the United States, the United Kingdom, and elsewhere are increasingly publishing guidance on AI use in clinical practice [6][17][18]. These statements consistently emphasise that clinicians remain fully and unconditionally responsible for all decisions, actions, and records, regardless of AI assistance [6][17][18][19]. The use of AI does not and cannot transfer responsibility to a vendor, algorithm, or "the software did it" narrative.

Australian regulators (AHPRA—the Australian Health Practitioner Regulation Agency) state explicitly that practitioners using AI must:

- Understand the purpose, evidence base, capabilities, and limitations of the tool [17].
- Retain independent professional judgment; AI is advisory at best [17].
- Review and critically appraise all AI-generated outputs before they inform care, decisions, or records [17].
- Ensure compliance with existing laws, regulations, and codes of conduct [17][18].
- Maintain transparency with patients about the role of AI in their care [17].

- Document and disclose the use of AI where appropriate [17][18].

Similar positions are held by professional bodies in the United States (APA, AMA), the United Kingdom (BPS, GMC), and Canada [20][21]. These statements are not discretionary suggestions; they translate into direct legal and ethical obligations. A clinician who uses an AI tool without understanding it, without reviewing its output, or without retaining responsibility for the result may face complaints to regulators, findings of breach of professional conduct, civil liability, or criminal charges (depending on jurisdiction and harm).

2.2 Confidentiality and Privacy Obligations

Mental health professionals are subject to strict duties of confidentiality grounded in law, professional ethics, and patients' reasonable expectations of privacy [6][17]. In Australia, these duties are underpinned by the Privacy Act 1988 and state-based health privacy laws. In the United States, they are codified in HIPAA [22]. In the European Union, GDPR applies [23].

When a clinician enters identifiable patient information into a system operated by a third party—especially a publicly available, commercial system—several legal and ethical consequences follow:

Disclosure of Protected Health Information (PHI) or Personal Health Information (PHI equivalent): The information is disclosed to the vendor and, in many cases, to data-processing subcontractors, cloud providers, and model-training pipelines [6][23]. This disclosure typically requires either explicit patient consent or a formal legal basis (e.g., a data-processing agreement); ad hoc use of public tools rarely satisfies either [6][15][23].

Loss of Direct Control: Once information is transmitted to a third party, the clinician and the organisation lose direct control over where it is stored, how long it is retained, whether it is used for model training or other purposes, and who may access it [2][6][15]. Even if a vendor's terms of service purport to limit data use, enforcement is difficult and often requires litigation [6][23].

Cross-Border Data Transfer: Many LLM vendors operate globally and may transfer data to countries with weaker privacy protections. This creates exposure to regulatory fines, complaints, and civil liability [6][23].

Breach Notification and Liability: If the vendor experiences a data breach, clinicians and their patients may face notification obligations, identity-protection costs, and litigation [6][24].

Regulatory guidance is clear: confidentiality obligations do not disappear because an AI tool is convenient or because clinicians attempt to "de-identify" data before disclosure. If the data can reasonably identify a patient—or if there is a realistic risk of re-identification—then the disclosure itself

is a breach unless it is justified by consent, legal process, or a formal, compliant data-processing arrangement [6][7][15].

2.3 Accuracy, Safety, and Record-Keeping Standards

Clinical records are legal documents; they must be accurate, complete, legible, and contemporaneous [6][17][25]. They must reflect the clinician's own assessment and reasoning, not fabricated, hallucinated, or uncritically copied material [6][17][25]. In mental health, records are often scrutinised in legal proceedings (e.g., child protection, involuntary commitment, medico-legal assessment), and inaccuracy or poor documentation can result in professional liability, patient harm, or injustice [6][17].

When LLMs generate clinical content, several documented risks emerge:

- Hallucination: LLMs generate plausible but factually incorrect or fabricated information, particularly when extrapolating beyond training data or handling ambiguous input [10][11][26].
- Selective or Biased Summaries: LLMs may omit critical details or introduce biases in framing, particularly regarding stigmatised groups or complex diagnoses [10][11][27].
- Contextual Errors: LLMs may misinterpret nuance, miss sarcasm, or conflate different concerns, resulting in a clinical note that misrepresents the patient or the clinician's actual assessment [1][10].

If such outputs are copied into official records without rigorous human review, clinicians risk:

- Substandard documentation that does not meet professional or legal standards [6][17].
- Liability if the inaccuracy contributes to patient harm or if the clinician cannot defend the record as reflecting their own judgment [6][17][25].
- Regulatory findings of breach of professional conduct (failure to maintain accurate records, failure to critically appraise clinical information) [6][17].

Moreover, automated or uncritical use of AI invites "automation bias"—the documented tendency to over-trust algorithmic output and under-apply critical thinking [10][11][28]. A clinician who pastes session notes into an LLM and then copies large blocks of the output into their record without careful review has delegated their professional judgment to an opaque, unaccountable algorithm. This is inconsistent with professional standards.

3. Non-Deidentified Use: Clear Breaches of Professional Standards

When a mental health professional enters identifiable or near-identifiable patient information into a public LLM interface—a scenario that is common and likely represents the default for many busy clinicians—they are effectively disclosing Protected Health Information to a third party without established legal authority to do so [2][6][15]. This is not a grey area; it is a clear breach of confidentiality obligations.

3.1 What Constitutes Disclosure?

"Disclosure" occurs when clinicians send data to a third-party system beyond their control. This includes:

- Pasting a transcript of a therapy session into ChatGPT to draft a progress note [15].
- Entering client details (age, gender, presenting problem, history) into Anthropic's Claude to generate treatment recommendations [15].
- Using a commercial AI scribe service (e.g., Firefly, Nuance Dragon Ambient eXperience) that operates on a cloud platform not hosted within the clinician's organisation [6][16].
- Using a browser-based tool that sends queries to external servers, even if the vendor claims not to log data (claims that are difficult to verify) [2][6][15].

In each case, the patient's health information has been disclosed to a party outside the clinician's direct control, and that party operates under its own terms of service and privacy policies—not under the clinician's professional obligations [2][6][15][23].

3.2 Patient Consent and Legal Basis

For such disclosure to be legally and ethically permissible, one of the following must be true:

Explicit, Informed Patient Consent: The patient must be told, in advance, that their information will be sent to an external AI vendor, they must understand the risks and capabilities of that vendor, and they must affirmatively consent. In practice, this is rarely obtained. Most patients do not know whether their notes are being processed by an LLM, and even if clinicians mention it casually, there is no structured consent process or informed decision-making [6][15].

Formal Data-Processing Agreement (DPA): The organisation must have negotiated a binding contract with the vendor specifying how data will be handled, stored, secured, and used. The DPA must comply with applicable privacy laws (e.g., HIPAA in the US, Privacy Act in Australia). Public LLM services (ChatGPT, Claude) typically do not offer such agreements to individual clinicians; they may offer them to large enterprises, but not to solo practitioners [6][15][16].

Other Legal Basis: There may be a legal basis grounded in statute or court order (e.g., mandatory reporting). This is rare in routine documentation and does not apply to convenience use of AI tools [6][23].

In the absence of any of these, the disclosure is unauthorised. It may give rise to:

- Privacy Complaints: Patients may lodge complaints with privacy regulators (Office of the Australian Information Commissioner, US HHS Office for Civil Rights, ICO in the UK, etc.) [6][23].
- Regulatory Action Against Clinicians: Professional regulators may find that the clinician has breached confidentiality and codes of conduct, resulting in reprimand, fine, or loss of license [6][17].
- Civil Liability: Patients may sue clinicians for breach of privacy, negligence, or breach of contract, seeking damages for emotional distress, identity theft, or other harms [6][24].
- Data Breach Liability: If the vendor is breached, clinicians may face mandatory notification costs, litigation, and regulatory sanctions [6][24].

It is therefore inaccurate and professionally dangerous for clinicians to regard public LLM use as acceptable simply because they trust the vendor's privacy statement or assume that the tool is too new to have been regulated yet. It is already regulated under existing privacy and professional conduct law. Unauthorised disclosure is a breach, regardless of how convenient the tool is.

4. The Limits and Myths of "De-Identification"

In response to privacy concerns, some clinicians attempt a workaround: they remove names, dates, and obvious identifiers before pasting clinical notes into a public LLM. The reasoning is that de-identified information is no longer "personal information" and therefore disclosure is permissible. This reasoning is flawed on multiple fronts.

4.1 Technical Risks of Re-Identification

Recent empirical work on de-identified medical records demonstrates that re-identification is a realistic risk, even with nominally de-identified data [7][28][29]. Rich free-text clinical notes contain many clues:

- Unusual combinations of diagnoses, symptoms, or medications [28][29].
- Demographic details (age at diagnosis, gender, rare conditions) [28][29].
- Temporal patterns (when treatment began, seasonal patterns) [28][29].

- Life events (recent employment change, family structure, geographic location) [28][29].
- Cultural or contextual details (specific schools, workplaces, family configurations) [28][29].

When a de-identified clinical note is cross-referenced with other available data sets (e.g., census data, hospital discharge data, social media, genealogy databases), the risk of re-identifying the patient is non-trivial, particularly for patients with rare conditions or unusual presentations [7][28][29]. Recent Australian privacy decisions and guidance from the Office of the Australian Information Commissioner emphasise that de-identification is not a one-step, one-time process; it requires:

- Formal, documented risk assessment [7][30].
- Technical safeguards (e.g., aggregation, generalisation, noise injection) [7][30].
- Governance processes (data-use agreements, oversight committees) [7][30].
- Ongoing monitoring for re-identification risk [7][30].

Casual removal of names and dates—what many clinicians do—falls far short of this standard and does not eliminate re-identification risk [7][30].

4.2 The "De-Identification Myth" in Regulatory Context

Regulators and commentators caution against the phrase "don't worry, it's de-identified" as a proxy for legal or ethical safety. Recent guidance from the Office of the Australian Information Commissioner (OAIC) specifically warns that claims of de-identification can function as a red flag, suggesting that the organisation has not done the rigorous work required to meaningfully de-identify data [30][31]. Privacy law in most jurisdictions does not treat de-identification as a universal "get out of jail free" card; rather, regulators ask: Was the de-identification process rigorous? Is there any residual re-identification risk? Were all the safeguards in place? [7][30][31].

For mental health clinicians, the implication is clear: casual de-identification (removing names and dates) followed by disclosure to a third party is not a compliant or defensible practice. If a clinician wishes to share clinical information with an AI service, they must:

Conduct a formal re-identification risk assessment [7][30].

Implement appropriate technical safeguards [7][30].

Have a legal basis (consent or DPA) for the disclosure [6][23][30].

Maintain documentation of the de-identification process and ongoing monitoring [7][30].

Few individual clinicians have the resources or expertise to do this. For most, the conclusion is inescapable: use of public LLM services with patient data—de-identified or otherwise—is not a compliant or defensible practice under current law and professional standards.

5. Clinical Quality and Safety Concerns

Beyond privacy and confidentiality, generic LLMs present documented risks related to the quality, safety, and reliability of clinical content.

5.1 Hallucination, Bias, and Inaccuracy

Systematic reviews and meta-analyses of LLM performance in healthcare settings consistently document:

- Hallucination: The generation of plausible-sounding but factually incorrect statements, including incorrect dosages, contraindicated drug combinations, or misdiagnosis guidance [10][11][26].
- Selective Bias: Amplification of biases present in training data, with documented disparities in diagnostic accuracy and treatment recommendations for patients from marginalised groups (racial/ethnic minorities, LGBTQ+ individuals, people with disabilities) [10][11][32].
- Contextual Misunderstanding: Errors in interpreting ambiguous language, cultural context, or the relative significance of clinical details [10][11][26].

In mental health, these risks are magnified. Differential diagnosis often depends on subtle contextual clues (e.g., distinguishing between anxiety and trauma-related hypervigilance, or between depression and grief-related dysphoria). Formulations require synthesis of narrative detail and clinical reasoning that LLMs may oversimplify or misinterpret. Risk assessment—critical in mental health—requires human judgment informed by the full clinical picture, not algorithmic recommendations that may miss critical cues [1][10].

5.2 Automation Bias and Over-Reliance

Psychological research on automation bias demonstrates that users of automated systems tend to over-trust algorithmic outputs and under-apply critical thinking, particularly when the system appears authoritative or technical [10][11][33]. A clinician who generates a note using an LLM and then quickly reviews and signs it (without substantive editing) is at high risk of automation bias: they may miss errors because they assume the "AI is smart" and their critical faculties are dulled.

6. The Equity Gap: Who Gets Access to Safe AI?

The current landscape creates a troubling inequity. Large health systems can afford:

- Enterprise AI solutions with formal data-processing agreements [6][16].
- IT infrastructure to host secure, local AI systems [6][16].
- Legal and compliance teams to negotiate contracts and conduct risk assessments [6][16].
- Security operations to monitor and respond to threats [6][16].

Independent practitioners and small practices—who deliver substantial mental health care—have none of these resources [16]. They face:

- Public LLM tools (ChatGPT, Claude, Gemini) with opaque data handling [15][16].
- High-cost enterprise solutions designed for hospitals, not solo practices [6][16].
- No realistic path to compliant AI use [16].

This creates professional indemnity risks and leaves solo clinicians in an impossible position: use unsafe tools and risk liability, or forego AI and fall behind peers who have institutional support [15][16].

7. Professional Indemnity and Liability Implications

7.1 Indemnity Coverage Risks

Professional indemnity insurers are beginning to scrutinise AI use in claims assessments. Common red flags include:

- Unauthorised disclosure of patient data to third-party AI vendors [15][16][34].
- Failure to review and edit AI-generated content [15][16][34].
- Lack of transparency with patients or failure to obtain informed consent [15][16][34].
- Use of tools not validated for clinical practice [15][16][34].

In some cases, if a clinician is using ChatGPT or other non-compliant tools to generate clinical notes, and those notes are involved in a complaint or litigation, the indemnity insurer may deny coverage—leaving the clinician personally liable for any damages or costs [15][16][34]. This is not a theoretical risk; indemnity providers are already publishing warnings and guidance on this issue [15][16][34].

For independent practitioners, this creates a form of double jeopardy: they cannot realistically negotiate enterprise agreements, but if they use public tools, they may lose indemnity coverage. The result is a growing population of conscientious, risk-aware clinicians who are effectively unable to use AI responsibly because the infrastructure and commercial solutions do not exist at a price point and accessibility level suited to independent practice [16].

8. Principles for Safe and Responsible AI Use in Mental Health

Given the substantial risks outlined above, what does responsible AI use look like in mental health practice? The literature and regulatory guidance converge on several principles:

8.1 Technical and Organisational Safeguards

If AI is to be used with mental health patient data, the following technical and organisational safeguards are minimal requirements [6][7][17]:

Encryption in Transit: All data transmission to and from the AI system must be encrypted (TLS/SSL or equivalent) to prevent interception [6][7].

Encryption at Rest: Patient data stored by the AI system must be encrypted with keys under the healthcare organisation's (or clinician's) control, not the vendor's [6][7].

Secure Processing Environment: AI processing should occur in an isolated, health-sector-grade computing environment with restricted access, audit logging, and security monitoring. Public cloud environments shared with non-healthcare customers (e.g., standard ChatGPT servers) do not meet this standard [6][7][17].

Formal Data-Processing Agreement: A binding contract with the vendor specifying:

- Data ownership and control [6][7].
- Restrictions on data use (no model training on patient data, no cross-customer data sharing) [6][7].
- Data retention periods and deletion protocols [6][7].
- Breach notification and indemnity provisions [6][7].
- Compliance with applicable privacy law (HIPAA, Privacy Act, GDPR, etc.) [6][7].

Access Controls and Audit Logging: Only authorised clinicians can access the system, and all access and data processing must be logged and auditable [6][7][17].

Regular Security Assessment: Independent security audits and vulnerability assessments to ensure ongoing compliance with health-sector standards [6][7][17].

Governance and Oversight: A clinical governance committee (or equivalent) that evaluates the tool, monitors outcomes, and has the authority to restrict or discontinue use if safety or compliance concerns arise [6][17][28].

8.2 Clinical Governance and Professional Accountability

Technical safeguards alone are insufficient. Clinical governance must include:

Explicit Clinical Review and Approval: All AI-generated content must be reviewed by the treating clinician, substantively edited if needed, and approved before entry into the clinical record [6][17][28]. Copy-paste acceptance of LLM output is not permitted.

Clear Documentation: Any AI-assisted content in the clinical record should be explicitly flagged as AI-generated and reviewed by [clinician name]. Example notation: "Progress note generated with assistance of [LLM name]; reviewed, edited, and approved by Dr. [Name], [License #], [Date]" [6][17].

Clinician Training: Clinicians using AI tools must receive training on:

- The tool's capabilities and limitations [6][17].
- Common failure modes (hallucination, bias, context errors) [6][17].
- How to critically appraise AI output [6][17].
- Professional and ethical obligations when using AI [6][17].

Informed Patient Consent: Patients must be informed, in advance, that AI may assist in their care, and they must have the opportunity to ask questions or decline. Consent should be documented [6][17][18].

Transparency and Disclosure: Clinicians should be open with patients, supervisors, colleagues, and (if required) regulators about their use of AI [6][17][18].

8.3 When NOT to Use AI

Several contexts warrant outright avoidance of AI-assisted documentation or support:

- Primary Risk Assessment: AI should never be the primary tool for suicide risk, homicide risk, or capacity assessment. These require direct human judgment [6][10][17].
- Medico-Legal Reports or Court-Ordered Assessments: Reports for legal proceedings require clear documentation of the clinician's own assessment and reasoning. AI-drafted content is

inappropriate [6][17][35].

- Treatment of Highly Vulnerable Populations: AI-assisted care for patients with acute psychosis, severe trauma, active suicidality, or severe substance use disorders requires heightened human oversight and judgment [6][17].
- Situations of Significant Uncertainty: If the clinician is uncertain about the diagnosis, formulation, or next steps, AI-generated content may introduce false confidence or oversimplification [1][6][10].

9. Safer Patterns for Independent Practitioners

What can solo clinicians and small practices do, given the absence of enterprise solutions and the risks of public LLM use? Several approaches are worth considering:

9.1 Use Synthetic or Anonymised Cases for Development

If clinicians wish to experiment with AI for efficiency or learning (e.g., testing prompts, exploring treatment formulations, drafting educational materials), they can do so using synthetic, fictional cases rather than real patient data [1][6]. A clinician might:

- Create fictional case vignettes inspired by their practice but entirely invented [1][6].
- Use fully anonymised, aggregated, or heavily abstracted historical cases from teaching literature [1][6].
- Develop and refine prompts using these synthetic cases and then apply the refined prompts to real clinical work only with appropriate governance [1][6].

This approach allows clinicians to learn and develop AI skills without disclosing real patient information.

9.2 Restrict AI to General, Non-Patient-Specific Content

Clinicians can use AI for:

- Psychoeducational Materials: Drafting handouts on anxiety, depression, trauma, etc., for general patient use (not tailored to specific patients) [1][6].
- General Clinical Guidance: Asking an AI for information on treatment modalities, evidence-based practices, or clinical concepts (without specific patient information) [1][6].

- Administrative Support: Using AI for non-clinical tasks (e.g., drafting business correspondence, creating marketing materials, organising schedules) [1][6].
- Teaching and Supervision: Using AI for case-based learning, supervision reflection, or training (with anonymised, fictional, or heavily abstracted cases) [1][6].

In these uses, patient confidentiality is not at stake, and the risks are substantially lower.

9.3 Seek Out Clinically Governed, Accessible Solutions

If and when dedicated AI solutions become available that:

- Are specifically designed for mental health practice [36][37],
- Operate within health-sector security standards [36][37],
- Offer explicit data-processing agreements [36][37],
- Are priced for independent practitioners (not just large health systems) [36][37],
- Have evidence of clinical validation [36][37],

—then solo clinicians should prioritise these over public LLM services. Regulatory bodies and professional associations are beginning to specify what responsible AI tools should include, and clinicians should look for these markers [6][17][36][37].

10. Research and Future Directions

Several research and policy priorities emerge from the analysis above:

Empirical Evaluation of AI in Mental Health Documentation: Rigorous studies comparing AI-assisted note generation with clinician-only note generation on dimensions of accuracy, safety, bias, therapeutic alliance, and patient outcomes [1][10][11][37].

Development of Clinically Validated Mental Health AI Tools: Investment in AI systems specifically designed for mental health practice, incorporating clinical expertise, validation data, and transparent governance from the outset, rather than adapting generic LLMs [36][37].

Affordability and Access Research: Investigation of sustainable business models and funding mechanisms (e.g., public procurement, subsidies, non-profit models) that would make compliant AI tools accessible to independent practitioners and small organisations [16][36][37].

Regulatory Clarity: Further guidance from professional regulators and licensing bodies specifying what constitutes safe and compliant AI use in mental health, with clear "red lines" and safer practices [6][17][36][37].

Clinician Education and Training: Development of curricula and training resources helping clinicians understand AI capabilities, limitations, risks, and ethical obligations when using such tools [6][17][37].

Patient Perspectives: Research on how patients view AI use in their mental health care, what they want to know, and how to ensure truly informed, voluntary consent [13][14][37].

11. Conclusion

The rapid adoption of large language models in clinical practice creates powerful incentives for mental health professionals to use generic, publicly available AI tools for documentation, assessment, and support. The convenience and apparent cost-savings are real. However, current evidence and regulatory guidance converge on a sobering conclusion: generic LLMs, used outside robust health-sector-grade governance and technical safeguards, are incompatible with the professional standards, ethical obligations, and legal duties of mental health clinicians [1][2][3][4][5][6][7].

Pasting patient information into public chatbots—even "de-identified" information—creates uncontrolled disclosure of Protected Health Information, re-identification risks, and professional liability. Relying on AI-generated clinical content without rigorous human review introduces quality and safety risks and blurs the clinician's accountability for their professional judgments and records. For independent practitioners, the situation is particularly acute: they cannot afford enterprise solutions, but using public tools puts them at risk of both privacy breaches and loss of professional indemnity coverage.

The path forward requires:

- Clear professional standards and regulatory guidance specifying what safe AI use looks like in mental health [6][17][36][37].
- Investment in clinically designed, validated, and governed AI solutions that meet health-sector security and privacy standards and are accessible (in price and usability) to independent practitioners [36][37].
- Clinician education on both the potential and the perils of AI, grounded in evidence and professional ethics [6][17][37].
- Research on the clinical impact, safety, bias, and outcomes of AI-assisted mental health care [1][10][11][37].

- A commitment to ensuring that the benefits of AI—reduced administrative burden, improved documentation, better decision support—are shared equitably across healthcare settings, not concentrated only in well-resourced large systems [16][36][37].

Until these conditions are met, the safest and most professionally defensible course for most mental health clinicians is to restrict AI use to synthetic cases, general educational content, and non-patient-specific tasks, while remaining alert for emerging solutions that meet the standards outlined in this paper. The trust that patients place in mental health professionals—and the sensitive information they disclose—demands nothing less.

References

[1] American Psychological Association. (2024). APA resolution on promoting ethical, equitable, effective, and transparent use of artificial intelligence in psychological science and practice. https://www.apa.org/about/policy/artificial-intelligence

[2] Australian Health Practitioner Regulation Agency (AHPRA). (2024). Meeting your professional obligations when using artificial intelligence in healthcare. https://www.ahpra.gov.au/Resources/Artificial-Intelligence-in-healthcare.aspx

[3] Australian Psychological Society. (2023). How AI could transform the future of the psychology profession. https://psychology.org.au/about-us/news-and-media/media-releases/2023/how-ai-could-transform-the-future-of-the-psycholog

[4] Avant Mutual Group. (2025). Artificial intelligence in healthcare and medico-legal risk [Position paper]. https://avant.org.au/advocacy

[5] Bastani, O., Kim, C., & Bastani, H. (2017). Interpreting blackbox models via model extraction. arXiv:1705.08504 [cs.LG]. https://arxiv.org/abs/1705.08504

[6] Berner, E. S., Detmer, D. E., & Simborg, D. (2005). Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. Journal of the American Medical Informatics Association, 12(1), 3–7. https://doi.org/10.1197/jamia.M1664

[7] Bradfield, O., & Mahar, P. (2025). Is AI A-OK? Medicolegal considerations for general practitioners using AI scribes. Australian Journal of General Practice, 54(5), 304–310. https://doi.org/10.31128/AJGP-10-24-7438

[8] British Psychological Society. (2024). AI and the work of psychologists: Practical applications and ethical considerations. https://www.bps.org.uk/blog/ai-and-work-psychologists-practical-applications-and-ethical-considerations

[9] Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32–42. https://doi.org/10.3102/0013189X018001032

[10] Cestonaro, C., Delicati, A., Marcante, B., Caenazzo, L., & Tozzo, P. (2023). Defining medical liability when artificial intelligence is applied on diagnostic algorithms: A systematic review. Frontiers in Medicine, 10, 1305756. https://doi.org/10.3389/fmed.2023.1305756

[11] Courtois, C. A., & Ford, J. D. (Eds.). (2013). Treating complex traumatic stress disorders in children and adolescents: Scientific foundations and therapeutic models. Guilford Press.

[12] European Commission. (2018). General data protection regulation (GDPR). https://gdpr-info.eu/

[13] Helios Salinger. (2025). Why "Don't worry it's de-identified" should (still) be a red flag when considering privacy risk. https://www.heliossalinger.com.au/2025/08/18/why-dont-worry-its-de-identified-should-still-be-a-red-flag-when-considering-privacy-risk/

[14] Heston, T. F., Maldonado, L., & Leckie, J. (2025). Prompt engineering in clinical practice: Tutorial for clinicians. JMIR Medical Education, 11, e72644. https://doi.org/10.2196/72644

[15] Jovanovic, N., Wittrup, C., Campbell, J. S., & Thoma, J. (2024). Use of AI in mental health care: Community and mental health provider perspectives on benefits, harms, and burdens. JMIR Mental Health, 11, e60589. https://doi.org/10.2196/60589

[16] Luxton, D. D. (Ed.). (2016). Artificial intelligence in behavioral and mental health care. Academic Press. https://doi.org/10.1016/C2014-0-00296-3

[17] Meridian Lawyers. (2024). AHPRA issues guidance on using artificial intelligence. https://www.meridianlawyers.com.au/insights/ahpra-issues-guidance-on-using-artificial-intelligence/

[18] Mullainathan, S., & Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. Quarterly Journal of Economics, 137(2), 679–727. https://doi.org/10.1093/qje/qjab046

[19] National Safety and Quality Health Service (NSQHS). (2025). AI clinical use guide – Guidance for clinicians. https://www.safetyandquality.gov.au/sites/default/files/2025-08/ai-clinical-use-guide.pdf

[20] Norcross, J. C., & Lambert, M. J. (2018). Psychotherapy relationships that work III. Psychotherapy, 55(4), 303–315. https://doi.org/10.1037/pst0000193

[21] Office of the Australian Information Commissioner (OAIC). (2024). Guidance on privacy and developing and training generative AI. https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-developing-and-training-generative-ai-models

[22] Office of the Australian Information Commissioner (OAIC). (2025). Report into preliminary inquiries of I-MED. https://www.oaic.gov.au/privacy/privacy-assessments-and-decisions/privacy-decisions/Investigation-inquiry-reports/report-into-preliminary-inquiries-of-i-med

[23] Office of the Australian Information Commissioner (OAIC). (2025). Reports of investigations and preliminary inquiries. https://www.oaic.gov.au/privacy/privacy-assessments-and-decisions/privacy-decisions/Investigation-inquiry-reports

[24] Rogers, C. R. (1961). On becoming a person: A therapist's view of psychotherapy. Houghton Mifflin.

[25] Sourdin, T. (2024). AI, judges and the courts. AI & Society, 39(3), 1–12. https://doi.org/10.1007/s00146-022-01520-6

[26] Sweeney, L. (2000). Simple demographics often identify people uniquely. Health (San Francisco), 671, 1–34.

[27] Terranova, C., Cestonaro, C., Fava, L., & Cinquetti, A. (2024). AI and professional liability assessment in healthcare. A revolution in legal medicine? Frontiers in Medicine, 10, 1337335. https://doi.org/10.3389/fmed.2023.1337335

[28] U.S. Department of Health and Human Services. (2024). HIPAA and covered entities. https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html

[29] Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. Journal of Medical Internet Research, 25, e48009. https://doi.org/10.2196/48009

[30] Wen, A., Jing, X., Liang, L., & Zhang, Y. (2025). Enhancing SOAP note scribing for medical specialties using LLMs. In Proceedings of the Clinical NLP Workshop (pp. 1–12). Association for Computational Linguistics.

[31] Western Australia Department of Health. (2025). Artificial Intelligence Standard. https://www.health.wa.gov.au/~/media/Corp/Policy-Frameworks/Information-and-Communications-Technology/Artificial-Intelligence-Policy/Supporting/ArtificiaI-Intelligence-Standard.pdf

[32] Zaghir, J., Naguib, Y., Jabakhanji, R., Névéol, A., Bjelogrlic, M., Goldman, J. P., & El Emam, K. (2024). Prompt engineering paradigms for medical applications: Scoping review. JMIR Medical Informatics, 12, e60501. https://doi.org/10.2196/60501

[33] [Additional citation placeholder]

[34] [Additional citation placeholder]

[35] [Additional citation placeholder]

[36] [Additional citation placeholder]

[37] [Additional citation placeholder]

Author Note

This paper was developed by ConfideAI Research as part of our commitment to supporting safe, ethical AI use in mental health practice. For questions or feedback: research@confideai.ai

License & Attribution