Revista Ibero-Americana de Humanidades, Ciências e Educação — REASE

doi.org/10.51891/rease.v12i11.31556

INTELIGÊNCIA ARTIFICIAL GENERATIVA NA DOCUMENTAÇÃO PSIQUIÁTRICA: ACURÁCIA, ERROS E SUPERVISÃO HUMANA — REVISÃO NARRATIVA

GENERATIVE ARTIFICIAL INTELLIGENCE IN PSYCHIATRIC DOCUMENTATION: ACCURACY, ERRORS, AND HUMAN OVERSIGHT — A NARRATIVE REVIEW

INTELIGENCIA ARTIFICIAL GENERATIVA EN LA DOCUMENTACIÓN PSIQUIÁTRICA: PRECISIÓN, ERRORES Y SUPERVISIÓN HUMANA — REVISIÓN NARRATIVA

Letícia Kassai da Costa[1]
Lívia Baraldi Martelo[2]

RESUMO: Introdução: A inteligência artificial generativa pode apoiar a documentação psiquiátrica, mas a fluência textual pode ocultar falhas clínicas. Objetivo: Analisar criticamente a acurácia, os erros documentais e a supervisão humana no uso de modelos de linguagem em Psiquiatria e saúde mental. Métodos: Revisão narrativa crítica baseada em buscas bibliográficas estruturadas no PubMed/MEDLINE, na LILACS/BVS e na SciELO, complementadas pelo Elicit, até setembro de 2026; 15 estudos constituíram o núcleo analítico. Resultados: Os estudos apontaram ganhos potenciais de eficiência, organização e legibilidade, mas também conteúdo não sustentado, omissões, distorções de significado e erros de atribuição. O desempenho variou conforme tarefa, modelo e instruções. A revisão humana foi frequentemente recomendada, embora pouco avaliada quanto à sua efetividade na prevenção de erros. Discussão: Qualidade linguística e satisfação profissional não equivalem à fidelidade clínica demonstrada. Conclusão: Os achados favorecem o uso assistivo, com conferência profissional e políticas institucionais de segurança, transparência e governança, sem sustentar a adoção documental autônoma.

Palavras-chave: Inteligência artificial generativa. Registros médicos. Segurança do paciente.

ABSTRACT: Introduction: Generative artificial intelligence can support psychiatric documentation, but fluent outputs may conceal clinical errors. Objective: To critically examine accuracy, documentation errors, and human oversight in the use of language models in psychiatry and mental health. Methods: A critical narrative review based on structured searches in PubMed/MEDLINE, LILACS via the Virtual Health Library, and SciELO, supplemented by Elicit, through September 2026; 15 studies formed the analytic core. Results: Studies reported potential gains in efficiency, organization, and readability, alongside unsupported content, omissions, distortions of meaning, and attribution errors. Performance varied substantially across the clinical documentation tasks, language models, and generation instructions examined. Human review by qualified professionals was frequently recommended, but its effectiveness in preventing errors was seldom directly evaluated. Discussion: Linguistic quality and clinician satisfaction do not establish clinical fidelity. Conclusion: Available findings favor assistive use with professional verification and institutional policies addressing safety, transparency, and governance; they do not support autonomous clinical documentation.

Keywords: Generative artificial intelligence. Medical records. Patient safety.

[1] Graduada em Medicina. Centro Universitário de Adamantina (FAI), Adamantina, São Paulo, Brasil.
[2] Acadêmica do curso de Medicina. Centro Universitário de Adamantina (FAI), Adamantina, São Paulo, Brasil.

RESUMEN: Introducción: La inteligencia artificial generativa puede apoyar la documentación psiquiátrica, pero la fluidez textual puede ocultar errores clínicos. Objetivo: Analizar críticamente la precisión, los errores documentales y la supervisión humana en el uso de modelos de lenguaje en Psiquiatría y salud mental. Métodos: Revisión narrativa crítica basada en búsquedas bibliográficas estructuradas en PubMed/MEDLINE, LILACS mediante la Biblioteca Virtual en Salud y SciELO, complementadas con Elicit, hasta septiembre de 2026; 15 estudios conformaron el núcleo analítico. Resultados: Los estudios señalaron posibles mejoras de eficiencia, organización y legibilidad, pero también contenido sin respaldo, omisiones, distorsiones de significado y errores de atribución. El desempeño varió según la tarea, el modelo y las instrucciones. La revisión humana se recomendó con frecuencia, aunque su eficacia para prevenir errores fue poco evaluada directamente. Discusión: La calidad lingüística y la satisfacción profesional no equivalen a una fidelidad clínica demostrada. Conclusión: Los hallazgos favorecen el uso como herramienta de apoyo, con verificación profesional y políticas institucionales de seguridad, transparencia y gobernanza, sin respaldar la documentación clínica autónoma.

Palabras clave: Inteligencia artificial generativa. Registros médicos. Seguridad del paciente.

1 INTRODUÇÃO

A documentação clínica ocupa posição central na assistência psiquiátrica, pois registra a avaliação psicopatológica, a evolução, os riscos identificados, as decisões terapêuticas e a comunicação entre profissionais. Além de sustentar a continuidade do cuidado, o registro pode influenciar decisões subsequentes sobre tratamento, encaminhamento, internação e alta. A carga de trabalho associada à produção desses documentos e o interesse em tecnologias capazes de apoiá-la têm motivado a introdução de ferramentas de transcrição e geração automática na prática psiquiátrica, ainda com base empírica limitada (BUCKLEY; WANG; GOPALAN, 2025).

A adoção dessas ferramentas também não é neutra sobre o próprio registro: em atenção primária, notas produzidas com escribas de IA apresentaram maior documentação de sintomas neuropsiquiátricos, acompanhada de menor probabilidade de intervenção psiquiátrica registrada (CASTRO et al., 2026).

Nesse contexto, modelos de linguagem de grande escala (large language models; LLMs) e sistemas de inteligência artificial generativa passaram a ser investigados para gerar, resumir, transformar ou estruturar notas clínicas, sumários de alta, cartas médicas e outros documentos. Em Psiquiatria e saúde mental, essas aplicações incluem desde ferramentas ambientais que transformam o diálogo da consulta em documentação até sistemas capazes de reescrever registros para linguagem voltada ao paciente e plataformas que geram notas de psicoterapia (STANTON et al., 2025; MCCRUDDEN et al., 2026; KHARKO et al., 2024). Pesquisas com

profissionais também mostram que ferramentas generativas já vêm sendo utilizadas em tarefas documentais na prática cotidiana (JOHNSON et al., 2026; SCHWARZ et al., 2026).

Os potenciais benefícios incluem redução do tempo administrativo, padronização de registros, apoio à sumarização de grandes volumes de informação e melhora de aspectos comunicacionais de determinados documentos (STANTON et al., 2025; MCCRUDDEN et al., 2026; KHARKO et al., 2024). Entretanto, a produção de texto fluente e formalmente plausível não garante fidelidade clínica. Estudos comparativos já documentaram conteúdo não sustentado, alucinações, erros de conteúdo, mudanças de significado e distorções introduzidas durante a geração ou transformação de notas psiquiátricas (SCHWIEGER et al., 2024; OZKARA MENEKSEOGLU et al., 2026).

A adoção dessas ferramentas na prática documental tem avançado paralelamente a uma base empírica ainda limitada sobre seus efeitos. Surveys com profissionais de saúde mental registram tanto expectativa de ganho documental quanto preferência pelas aplicações administrativas: parte expressiva dos psiquiatras já relatava uso de ferramentas generativas em questões clínicas e esperava documentação mais eficiente (BLEASE; WORTHEN; TOROUS, 2024), e a geração de notas de caso foi apontada como o uso mais útil entre profissionais de saúde mental (GOLDIE et al., 2025). Esse descompasso entre adoção e verificação é o que torna necessária a síntese das evidências sobre acurácia e supervisão.

Essa questão é particularmente relevante em Psiquiatria, área em que nuances linguísticas e contextuais podem alterar a interpretação de sintomas, risco, dinâmica familiar e condutas. Além disso, o desempenho dos LLMs pode variar conforme o modelo, a versão, o prompt, os parâmetros de geração e a disponibilidade de informação contextual.

Mesmo em tarefas estruturadas de avaliação psicopatológica, modelos com desempenho global competitivo apresentaram perfis de erro diferentes dos observados em clínicos (LENZ et al., 2026). Padrões semelhantes já foram descritos quando modelos generativos foram aplicados a outros materiais assistenciais e de pesquisa em saúde mental, com omissão de informação clínica, acréscimos não solicitados e termos alucinados em resumos gerados (RICKMAN, 2025; BALT et al., 2025).

A literatura disponível também é metodologicamente heterogênea. Parte dos estudos compara diretamente outputs com documentos de referência; outros avaliam legibilidade, estilo, similaridade textual ou qualidade percebida; surveys e estudos qualitativos descrevem experiências profissionais; e trabalhos de implementação ou governança abordam supervisão

humana, privacidade e segurança. Assim, percentuais obtidos por métodos distintos não podem ser interpretados como uma única medida de acurácia ou erro.

2 OBJETIVO

Diante dessa heterogeneidade, esta revisão narrativa tem como objetivo discutir criticamente as evidências sobre acurácia, erros, qualidade documental, consistência e supervisão humana no uso de inteligência artificial generativa e modelos de linguagem de grande escala aplicados à documentação em Psiquiatria e saúde mental. A análise diferencia erros diretamente verificados, indicadores automatizados, percepções profissionais e experiências de implementação, evitando que medidas de naturezas distintas sejam consideradas equivalentes.

3 MÉTODOS

3.1 Tipo de revisão e origem do material

Trata-se de revisão narrativa crítica da literatura sobre o uso de inteligência artificial generativa e modelos de linguagem de grande escala na documentação em Psiquiatria e saúde mental. A síntese teve caráter temático e interpretativo, sem pretensão de esgotar todas as publicações disponíveis.

3.2 Busca e delimitação das publicações

As buscas bibliográficas foram concluídas em 12 de setembro de 2026 nas bases PubMed/MEDLINE, Literatura Latino-Americana e do Caribe em Ciências da Saúde (LILACS), via Biblioteca Virtual em Saúde (BVS), e SciELO. O Elicit foi empregado apenas como ferramenta complementar de identificação de publicações. Foram consideradas publicações de janeiro de 2021 a setembro de 2026, em português, inglês e espanhol, utilizando termos relativos a inteligência artificial generativa, modelos de linguagem, Psiquiatria, saúde mental e documentação clínica.

Foram priorizados estudos revisados por pares que examinaram a geração, sumarização, transformação ou estruturação de documentos clínicos em psiquiatria e saúde mental, incluindo registros elaborados por escribas ambientais, e que abordaram qualidade, fidelidade, erros, segurança ou supervisão humana. Foram excluídos estudos sem aplicação documental,

pesquisas exclusivamente preditivas ou diagnósticas e relatos limitados a expectativas de uso futuro. A leitura integral concentrou-se em textos completos disponíveis gratuitamente.

Após a leitura dos títulos, resumos e textos completos, 15 estudos compuseram o núcleo de análise. Outras referências foram utilizadas apenas para contextualização e discussão. A seleção de textos completos gratuitos e a abrangência das bases consultadas são consideradas nas limitações da revisão.

3.3 Análise crítica e organização temática

Em cada estudo, foram examinados o contexto, o tipo de documento, a tecnologia utilizada, as formas de avaliação, os resultados e as limitações. A análise distinguiu acurácia e erros diretamente avaliados, indicadores automatizados, percepções profissionais e experiências de implementação. O Quadro 1 reúne as principais contribuições dos 15 estudos e a natureza das evidências, sem constituir classificação formal de qualidade metodológica.

Os achados foram organizados em quatro eixos: desempenho e qualidade documental; erros, omissões e distorções; consistência, reprodutibilidade e segurança; e supervisão humana. Diante da heterogeneidade dos estudos, não se estimou taxa global de erro nem se agregaram quantitativamente resultados obtidos por métodos distintos.

4 RESULTADOS

4.1 Panorama da literatura examinada

Os 15 estudos foram publicados entre 2024 e 2026, sendo três em 2024, um em 2025 e 11 em 2026, e abrangeram contextos da Alemanha, Estados Unidos, Reino Unido, Índia, Austrália, Suíça, China, Coreia do Sul, Emirados Árabes Unidos e Japão. Os delineamentos foram heterogêneos: seis estudos combinaram componentes qualitativos e quantitativos ou foram apresentados como métodos mistos, dois foram predominantemente qualitativos, seis corresponderam a benchmarks, comparações quantitativas ou análises computacionais/documentais, e um consistiu em estudo secundário de desenvolvimento de framework.

Os materiais analisados incluíram documentação e encontros clínicos reais, sessões de aconselhamento, entrevistas e surveys com profissionais, cenários simulados, pacientes padronizados, documentos institucionais e notas clínicas sintéticas. Foram avaliados modelos de linguagem generalistas e especializados, além de sistemas comerciais integrados a fluxos de

documentação clínica. Dois estudos (KHARKO et al., 2024; BOWDEN et al., 2026) incluíram corpus apenas parcialmente psiquiátrico, com resultados apresentados de forma agregada. As contribuições dos estudos são sintetizadas no Quadro 1.

Quadro 1 – Síntese das contribuições dos estudos examinados

Referência / ano	Principal contribuição	Natureza da evidência
LENZ et al. (2026)	Alguns LLMs alcançaram desempenho competitivo na avaliação psicopatológica estruturada; perfis de erro diferiram dos clínicos; supervisão avaliada apenas em simulação pós-hoc.	Acurácia diretamente avaliada + supervisão simulada
ADHIKARY et al. (2024)	Modelos especializados apresentaram bom desempenho de sumarização, mas a avaliação especializada identificou alucinações e limitações de confiabilidade para aplicação clínica.	Qualidade e erros diretamente avaliados
SCHWARZ et al. (2026)	Psiquiatras relataram uso documental, benefícios percebidos e problemas como conteúdo fabricado e necessidade de correção.	Percepções e experiências profissionais
JOHNSON et al. (2026)	Profissionais relataram transcrições inexatas, adições, omissões, distorções e referências inventadas, além de preocupações com governança.	Percepções profissionais/implementação
STANTON et al. (2025)	Tecnologia ambiental esteve associada a menor tempo administrativo e alta satisfação percebida, sem auditoria independente da precisão.	Implementação + qualidade percebida
SCHWIEGER et al. (2024)	Documentos humanos tiveram melhor avaliação global; alucinações foram verificadas em 40% dos 20 sumários de IA.	Erro e qualidade diretamente avaliados
KHARKO et al. (2024)	IA aumentou marcadores de empatia, mas textos ficaram mais longos, exigiram maior proficiência de leitura e apresentaram perdas de fidelidade.	Evidência indireta: qualidade/fidelidade em corpus clínico parcialmente psiquiátrico
LUO et al. (2026)	GPT-5 e Grok-4 melhoraram legibilidade, conteúdo e compreensão de consentimentos; não avaliou prontuários clínicos.	Qualidade documental objetiva
BOWDEN et al. (2026)	Prompt empático aumentou linguagem empática, mas o efeito foi predominantemente formulaico e dependente da instrução.	Evidência indireta: efeito do prompt/qualidade comunicacional; resultados agregados
YANG; HA (2026)	Regras automatizadas sinalizaram possíveis riscos; os autores distinguem proxies de erros confirmados; propõe framework human-governed.	Proxy automatizado + governança proposta
BOKHARI et al. (2026)	Psiquiatras e pacientes padronizados perceberam benefícios, mas relataram inferências inadequadas e necessidade de revisão humana.	Evidência qualitativa de implementação
OZKARA MENEKSEOGLU et al. (2026)	Identificou cinco categorias de erros: interpretação clínica, atribuição, distorção, abreviação/terminologia e estrutura/sintaxe.	Erro diretamente verificado
MCCRUDDEN et al. (2026)	Implementação em larga escala com revisão e pelo menos uma edição obrigatórias; feedback positivo não equivale a acurácia.	Implementação real + human-in-the-loop
GARG et al. (2026)	Notas sintéticas foram mais estruturadas, porém menos diversas lexical e semanticamente que corpora clínicos comparadores.	Qualidade estrutural de documentação sintética
ARIHISA et al. (2026)	Notas de LLMs foram mais longas, repetitivas e menos diversas; achados caracterizam estilo, não segurança clínica direta.	Consistência/estilo documental

Fonte: Elaboração própria com base nas publicações examinadas no manuscrito, 2026.

Nota: A natureza da evidência distingue acurácia e erros diretamente avaliados, métricas automatizadas, percepção de profissionais e estudos de implementação. O quadro não constitui classificação formal de qualidade metodológica.

Os achados são apresentados nos quatro eixos temáticos descritos a seguir.

4.2 Desempenho e qualidade documental

Os estudos que avaliaram diretamente o desempenho documental apresentaram resultados heterogêneos. Na comparação de sumários de alta psiquiátrica, documentos escritos por profissionais receberam avaliação global superior aos gerados pelo ChatGPT-4 (médias 3,78 e 3,12, respectivamente), com vantagem dos documentos humanos em 12 dos 15 domínios avaliados e em 16 dos 20 casos (SCHWIEGER et al., 2024).

Em sumarização de sessões de aconselhamento, um benchmark com 11 modelos identificou desempenho favorável de modelos adaptados ao domínio, incluindo MentalLlama, Mistral e MentalBART. Entretanto, as métricas automáticas não reproduziram integralmente a avaliação dos especialistas, que também identificaram limitações e alucinações nos resumos gerados (ADHIKARY et al., 2024).

Nos estudos de transformação documental, melhorias na apresentação do texto não corresponderam necessariamente a maior fidelidade. Em evidência indireta proveniente de corpus apenas parcialmente psiquiátrico, GPT-3.5 e GPT-4 produziram notas com mais marcadores de empatia e menor uso de abreviações não explicadas, porém mais extensas, com maior exigência de leitura e variações na fidelidade médica (KHARKO et al., 2024). Em 33 formulários de consentimento para aconselhamento, GPT-5 e Grok-4 melhoraram medidas de legibilidade, qualidade do conteúdo e compreensão em relação aos documentos originais (LUO et al., 2026). Também como evidência indireta, a comparação entre prompts neutro e empático em casos pediátricos mostrou maiores escores de empatia nas notas geradas sob instrução explícita, embora a linguagem permanecesse predominantemente genérica e dependente do prompt (BOWDEN et al., 2026).

Em tarefa estruturada de avaliação psicopatológica, GPT-5.1 e Gemini-3-Pro-Preview alcançaram acurácia global de 0,72 na condição de melhor desempenho. Para o GPT-5.1, as acurácias foram de 0,81 em depressão, 0,76 em mania e 0,60 em esquizofrenia, frente a médias clínicas de 0,79, 0,68 e 0,58, respectivamente. A comparação, contudo, envolveu diferentes

modalidades de entrada, com transcrições textuais para os modelos e registros audiovisuais para os profissionais (LENZ et al., 2026).

Nos estudos de implementação clínica, as avaliações dos profissionais foram predominantemente favoráveis, embora baseadas sobretudo em percepção. No piloto Anathem, aumentaram as avaliações de qualidade e acurácia percebidas com a documentação assistida por IA (STANTON et al., 2025). No Smart Notes, 97,7% das avaliações de profissionais em tempo integral e 98,4% das avaliações de profissionais contratados foram positivas; entretanto, essas avaliações foram voluntárias e não constituíram auditoria independente da acurácia documental (MCCRUDDEN et al., 2026).

4.3 Erros, omissões, alucinações e distorções

A caracterização mais detalhada dos erros foi apresentada em uma análise de notas psiquiátricas reais transformadas por GPT-3.5-Turbo, que identificou cinco categorias: má interpretação clínica, erros de atribuição, distorção de conteúdo, erros de abreviação ou terminologia e erros estruturais ou sintáticos (OZKARA MENEKSEOGLU et al., 2026). Em uma subamostra de 20 notas, foram identificados 134 erros, média de 6,7 por nota, sendo a especulação o subtipo mais frequente, com 55 ocorrências e presença em 95% das notas analisadas. Esses valores foram apresentados pelos autores como descrição da amostra, e não como estimativas populacionais. Entre os exemplos clinicamente relevantes estavam alterações no significado de conceitos psiquiátricos, inclusive suicidabilidade, atribuição incorreta de falas ou comportamentos e expansão inadequada de abreviações médicas.

Em sumários de alta psiquiátrica gerados pelo ChatGPT-4, alucinações foram identificadas em 40% dos 20 documentos, das quais 37,5% foram consideradas de elevada relevância clínica; erros menores de conteúdo ocorreram em 30% dos documentos de IA e em 10% dos documentos humanos (SCHWIEGER et al., 2024). Em resumos de sessões de aconselhamento, especialistas também identificaram diferentes graus de alucinação nos modelos avaliados (ADHIKARY et al., 2024).

Na avaliação psicopatológica estruturada, o GPT-5.1 apresentou 84 erros, enquanto as classificações modais dos clínicos apresentaram 69. Os perfis foram distintos: o modelo classificou predominantemente como “não avaliáveis” itens considerados ausentes pela referência, enquanto os clínicos tenderam a classificar como ausentes ou presentes itens considerados não avaliáveis. Essa diferença foi mais pronunciada nos itens dependentes de

observação, dos quais 63% foram classificados pelo modelo como não avaliáveis, frente a 0% no consenso clínico (LENZ et al., 2026).

Quanto às omissões, os resultados variaram entre os estudos. Na análise de notas psiquiátricas transformadas por GPT-3.5-Turbo, não foi identificada categoria específica de omissão (OZKARA MENEKSEOGLU et al., 2026). Em estudo com corpus apenas parcialmente psiquiátrico, foram descritas perdas de informação com repercussão sobre a fidelidade médica (KHARKO et al., 2024), consideradas evidência indireta nesta revisão. Em uma análise automatizada utilizada no desenvolvimento de um framework de segurança, sinalizadores de possível omissão foram observados em 2,9% dos resumos para a seção de fatores de risco, 8,6% para fatores de melhora e 13,7% para intervenções. Sinalizadores de omissão de conteúdo de alto risco ocorreram em 5,8% dos casos para autolesão ou suicídio, 15,1% para trauma ou violência e 13,7% para risco relacionado à dependência (YANG; HA, 2026).

No mesmo estudo, regras automatizadas sinalizaram possível conteúdo não sustentado em 41% dos 139 resumos, linguagem sobrediagnóstica em 31,7% e linguagem medicalizada em 54,7%; ao menos um sinalizador foi acionado em 91,4% dos casos. Esses resultados representam indicadores conservadores para revisão e não taxas confirmadas de alucinação, erro clínico ou insegurança (YANG; HA, 2026).

Nos estudos baseados em experiência profissional, os problemas foram identificados principalmente por relatos dos participantes. Foram descritos transcrições inexatas, acréscimos, omissões, distorções, referências inexistentes e necessidade de correções (JOHNSON et al., 2026; SCHWARZ et al., 2026). Estudos de escribas ambientais também relataram fabricações, inexatidões e inferências mais específicas do que as informações efetivamente exploradas durante a consulta (STANTON et al., 2025; BOKHARI et al., 2026).

4.4 Consistência, reprodutibilidade e segurança

O desempenho dos LLMs variou entre modelos e condições de geração. Na avaliação psicopatológica estruturada, GPT-5.1 e Gemini-3-Pro-Preview apresentaram os melhores resultados entre os modelos testados, enquanto alguns modelos mostraram maior sensibilidade às configurações utilizadas. Para a análise principal, o resultado do GPT-5.1 foi obtido por votação majoritária entre três execuções independentes (LENZ et al., 2026).

A geração repetida também integrou o desenho de outro estudo: para cada caso, foram produzidos três sumários de alta pelo ChatGPT-4 e a versão considerada melhor pelo pesquisador principal foi selecionada para avaliação (SCHWIEGER et al., 2024).

A dependência das instruções fornecidas ao modelo foi observada em diferentes contextos. Como evidência indireta, notas produzidas sob prompt explicitamente empático apresentaram escores de empatia superiores aos obtidos com prompt neutro, embora a linguagem permanecesse predominantemente genérica e dependente da instrução (BOWDEN et al., 2026). Em geração de notas clínicas sintéticas, prompts curtos e diretivos produziram maior coerência e fidelidade clínica do que instruções longas e descritivas, e foram utilizadas restrições explícitas para reduzir inferências não sustentadas (GARG et al., 2026). Em outra comparação, notas geradas por LLMs foram mais longas, repetitivas e lexicalmente menos diversas que as produzidas por médicos, com padrão de escrita mais uniforme e semelhante a templates (ARIHISA et al., 2026).

A reprodutibilidade temporal foi pouco explorada. Um estudo com corpus apenas parcialmente psiquiátrico repetiu a geração dos mesmos documentos em dois momentos separados por um mês, motivado pela possibilidade de variação das respostas ao longo do tempo, sem realizar, contudo, uma análise formal específica de reprodutibilidade temporal (KHARKO et al., 2024). Em outro estudo, realizado com dois modelos comerciais fechados, os autores destacaram que atualizações dos sistemas e o uso de interfaces web sem parâmetros fixos podem limitar a reprodução exata das gerações e produzir variação relacionada ao model drift (LUO et al., 2026).

Essas questões também apareceram nas propostas de governança. Um framework para uso de IA em saúde mental incorporou controle de versão, validação local, monitoramento pós-implementação, avaliação de drift e revalidação periódica como requisitos para níveis mais avançados de utilização clínica supervisionada (YANG; HA, 2026).

As métricas computacionais de similaridade e estrutura não foram tratadas como equivalentes à acurácia clínica. No estudo comparativo de estilos documentais, medidas como BLEU, ROUGE-L, BERTScore, TER e indicadores de redundância foram utilizadas para caracterizar padrões linguísticos e documentais; os próprios autores ressaltaram que essas análises não avaliavam acurácia diagnóstica ou correção clínica substantiva (ARIHISA et al., 2026). No Smart Notes, a similaridade textual pelo coeficiente Dice-Sørenson foi utilizada durante o desenvolvimento para orientar o refinamento dos prompts, enquanto o fluxo clínico

final exigia revisão e edição da nota pelo profissional antes da submissão (MCCRUDDEN et al., 2026).

4.5 Supervisão humana e limites para aplicação clínica

A necessidade de supervisão humana foi recorrente nos estudos, embora sustentada por diferentes tipos de evidência. O exemplo mais concreto de revisão humana incorporada ao fluxo documental foi o Smart Notes: a nota gerada era apresentada ao profissional, que deveria revisá-la e realizar pelo menos uma edição antes da submissão. Quando nenhuma modificação era detectada, o sistema impedia a finalização até que houvesse intervenção do profissional (MCCRUDDEN et al., 2026). O estudo demonstrou a operacionalização desse controle em uso clínico, mas não avaliou isoladamente quanto a revisão humana reduziu erros documentais.

No piloto Anathem, participantes manifestaram preocupação com complacência da automação e destacaram a necessidade de supervisão profissional e manutenção da responsabilidade clínica. Também foi levantada a possibilidade de perda de habilidades relacionadas à redação de documentos e à formulação psiquiátrica com o uso continuado da tecnologia (STANTON et al., 2025).

Em uma avaliação qualitativa de escriba ambiental realizada em ambiente simulado, calibração da confiança, consentimento, privacidade, treinamento e supervisão humana emergiram como temas relevantes para implementação. Um participante estimou que aproximadamente 90% do trabalho documental era realizado pela IA e 10% exigia edição. Também foram relatadas alucinações documentais e preocupação com excesso de confiança; um participante considerou que a própria imperfeição do sistema poderia reduzir o risco de dependência excessiva. Durante a simulação, apenas dois dos oito clínicos relataram ter solicitado permissão para gravação, evidenciando variabilidade na prática de consentimento naquele contexto experimental (BOKHARI et al., 2026).

A supervisão também foi explorada em simulação post hoc. Quando pares de clínicos discordavam na avaliação psicopatológica, a resolução por um psiquiatra certificado ou pelo LLM de melhor desempenho apresentou maior acurácia do que a escolha aleatória entre as duas classificações discordantes. O benefício foi observado nas três condições simuladas, mas não correspondeu à avaliação prospectiva de supervisão por IA em prática clínica real (LENZ et al., 2026).

Nos estudos que avaliaram diretamente erros ou qualidade documental, a revisão humana foi recomendada como salvaguarda. Na transformação de notas psiquiátricas, os autores destacaram a necessidade de revisão e aprovação profissional diante de erros capazes de alterar significado clínico (OZKARA MENEKSEOGLU et al., 2026). Os sumários de alta gerados pelo ChatGPT-4 foram propostos como rascunhos ou modelos para finalização por médicos (SCHWIEGER et al., 2024), enquanto os modelos avaliados na sumarização de sessões de aconselhamento foram considerados ainda insuficientemente confiáveis para aplicação clínica sem refinamento e validação adicionais (ADHIKARY et al., 2024).

Por fim, um estudo de desenvolvimento propôs um framework human-governed composto por sete etapas: entrada e desidentificação dos dados, execução da IA, barreiras estruturadas de segurança, revisão clínica, integração ao prontuário, monitoramento pós-implementação e governança institucional. As saídas da IA são mantidas como rascunho ou apoio à decisão, com controles direcionados a conteúdo não sustentado, omissões de risco, linguagem sobrediagnóstica, privacidade e completude (YANG; HA, 2026). O próprio estudo ressalta que o framework não estabelece, por si só, segurança ou efetividade clínica e ainda requer avaliação prospectiva, testes de usabilidade e validação independente antes da adoção rotineira.

5 DISCUSSÃO

5.1 Fidelidade clínica, comparabilidade e natureza da evidência

Esta revisão narrativa identificou que a inteligência artificial generativa vem sendo aplicada a diferentes tarefas de documentação em Psiquiatria e saúde mental, incluindo geração e transformação de notas clínicas, sumarização de sessões, elaboração de sumários de alta, produção de registros estruturados e documentação assistida por sistemas ambientais.

Entretanto, o conjunto das evidências não permite classificar esses sistemas, de forma geral, como acurados ou seguros. O desempenho variou conforme o modelo, o tipo de documento, as instruções fornecidas, a natureza dos dados e a estratégia utilizada para avaliação.

Um padrão recorrente foi o descompasso entre qualidade formal do texto e fidelidade clínica. Em documentação psiquiátrica, transformações por LLMs puderam melhorar aspectos linguísticos ou comunicacionais e, simultaneamente, introduzir interpretações inadequadas, atribuições incorretas, distorções e conteúdo especulativo (OZKARA MENEKSEOGLU et al., 2026).

Padrão semelhante foi observado em um estudo com corpus apenas parcialmente psiquiátrico, no qual melhorias linguísticas coexistiram com perda variável de fidelidade, resultado considerado evidência indireta nesta revisão (KHARKO et al., 2024). Esses achados indicam que fluência, legibilidade ou aparência formal do documento não podem ser utilizadas isoladamente como substitutos de acurácia clínica.

Esse fenômeno também foi descrito fora do corpus incluído nesta revisão. Em estudo de 100 atendimentos de emergência, 42% dos sumários produzidos pelo GPT-4 apresentaram alucinações e 47% omitiram informação clinicamente relevante, embora o potencial médio de dano atribuído aos erros tenha sido baixo (WILLIAMS et al., 2025). A convergência com os achados psiquiátricos sugere que a dissociação entre qualidade textual e fidelidade documental não se restringe à saúde mental e reforça a importância da conferência do texto gerado em relação à fonte clínica.

Essa distinção assume particular relevância em Psiquiatria, na qual termos relacionados a suicidabilidade, psicose, afeto, comportamento, relações familiares e resposta ao tratamento podem carregar significado clínico específico. Reformulações linguisticamente plausíveis alteraram conceitos psiquiátricos e o significado de comportamentos clinicamente relevantes (OZKARA MENEKSEOGLU et al., 2026). Além disso, mesmo em modelos com desempenho global competitivo, itens dependentes de observação permaneceram mais difíceis quando o modelo teve acesso apenas à transcrição textual da entrevista (LENZ et al., 2026).

Os estudos também mostraram que o termo “erro” representa fenômenos distintos conforme o método empregado. Comparações diretas entre fonte e documento produzido permitem identificar erros efetivamente presentes no texto (SCHWIEGER et al., 2024; OZKARA MENEKSEOGLU et al., 2026; ADHIKARY et al., 2024), enquanto regras automatizadas funcionam como sinalizadores de possíveis problemas e não como eventos clinicamente adjudicados (YANG; HA, 2026). Em surveys e estudos qualitativos, alucinações, omissões e distorções correspondem a experiências relatadas pelos profissionais (JOHNSON et al., 2026; SCHWARZ et al., 2026; BOKHARI et al., 2026). Essa distinção é necessária para evitar que resultados obtidos por métodos diferentes sejam interpretados como estimativas comparáveis de uma única taxa de falha.

A variabilidade entre modelos e condições de geração acrescenta outra limitação à comparabilidade das evidências. Versão do modelo, parâmetros, prompt, data e ambiente de execução podem influenciar os documentos produzidos (LENZ et al., 2026; LUO et al., 2026;

ARIHISA et al., 2026). Como evidência indireta, também foi demonstrado que uma modificação explícita no prompt alterou características comunicacionais das notas (BOWDEN et al., 2026).

Estratégias de geração repetida foram empregadas em alguns estudos, incluindo seleção do melhor entre três outputs (SCHWIEGER et al., 2024) e votação majoritária entre três execuções independentes (LENZ et al., 2026), condições que diferem da utilização de uma única geração sem seleção posterior. Paralelamente, análises linguísticas identificaram maior homogeneidade, repetição e menor diversidade em determinados documentos produzidos por LLMs (GARG et al., 2026; ARIHISA et al., 2026).

A interpretação exige atenção ao desenho e à fonte de cada resultado. Comparações diretas de documentos contra registros clínicos ou referências especializadas são mais informativas sobre erros efetivamente presentes do que estudos centrados em percepção de usuários. Trabalhos computacionais e de desenvolvimento de frameworks esclarecem propriedades textuais e controles propostos, mas não demonstram, isoladamente, segurança clínica. As publicações com corpus apenas parcialmente psiquiátrico e resultados agregados (KHARKO et al., 2024; BOWDEN et al., 2026) foram usadas somente como evidência indireta para fenômenos documentais, e não como estimativas específicas de acurácia na Psiquiatria.

5.2 Supervisão humana: recorrente nas recomendações, mas pouco testada

A supervisão humana foi um componente recorrente tanto nos estudos que identificaram erros quanto nos trabalhos de implementação e governança. Entretanto, as evidências disponíveis sustentam com maior clareza a necessidade de manter revisão profissional do que a eficácia de uma estratégia específica de supervisão.

O Smart Notes constitui o exemplo mais concreto de revisão humana incorporada ao fluxo assistencial: a nota gerada precisava ser revisada e receber pelo menos uma edição antes da submissão (MCCRUDDEN et al., 2026). Esse desenho demonstra que uma barreira obrigatória de intervenção profissional pode ser operacionalizada em larga escala, mas o estudo não mensurou quantos erros foram identificados ou corrigidos durante essa etapa. No estudo de Lenz et al. (2026), possíveis benefícios de supervisão foram observados apenas em uma simulação post hoc de resolução de discordâncias, enquanto o framework proposto por Yang MA e Ha KS (2026) ainda não foi validado prospectivamente em prática clínica.

À luz das evidências disponíveis, o uso com maior respaldo é assistivo, com produção de rascunhos submetidos à revisão e aprovação profissional antes da incorporação ao prontuário. Os tipos de erro identificados indicam que essa revisão não deveria se restringir à correção linguística ou formal, mas incluir conferência de conteúdo de risco, atribuições, sintomas, medicação, diagnóstico e plano terapêutico em relação às informações originais (SCHWIEGER et al., 2024; OZKARA MENEKSEOGLU et al., 2026; YANG; HA, 2026).

A possibilidade de complacência da automação constitui preocupação adicional. Em literatura anterior sobre sistemas clínicos de apoio à decisão, o fenômeno foi associado à verificação insuficiente de saídas automatizadas, sob influência da confiança no sistema, da carga de trabalho, da complexidade da tarefa e de restrições de tempo (GODDARD; ROUDSARI; WYATT, 2012). Na documentação psiquiátrica, esse risco pode favorecer a incorporação acrítica de informações imprecisas ao prontuário, com repercussões na continuidade assistencial e na segurança do paciente.

Treinamento, responsabilização dos usuários e características do desenho da interface figuram entre as medidas estudadas para reduzir o viés de automação (GODDARD; ROUDSARI; WYATT, 2012). No Brasil, a Resolução CFM nº 2.454/2026 estabelece a IA como apoio à prática médica, exige julgamento crítico do profissional e mantém sua responsabilidade ético-profissional pelos atos médicos que praticar com auxílio desses sistemas (CONSELHO FEDERAL DE MEDICINA, 2026). A norma também contempla transparência ao paciente, preservação da confidencialidade e comunicação de falhas relevantes; isso não equivale a imputar automaticamente ao médico falhas técnicas atribuíveis exclusivamente ao sistema. Em complemento, a Lei Geral de Proteção de Dados Pessoais considera os dados referentes à saúde como dados pessoais sensíveis e exige condições legais e medidas adequadas de segurança para seu tratamento (BRASIL, 2018). Tais dispositivos brasileiros fundamentam a necessidade de políticas institucionais, fluxos de verificação e delimitação de responsabilidades, mas não demonstram, por si sós, a efetividade de um modelo específico de revisão humana.

5.3 Implicações para a prática clínica

Os achados indicam que a implementação de ferramentas generativas para documentação deve considerar rastreabilidade, treinamento, consentimento, privacidade, revisão clínica e definição de responsabilidades. A conferência do documento com a fonte original (entrevista, transcrição ou registro clínico) é particularmente relevante para a

identificação de acréscimos não sustentados, omissões e alterações de significado. Entre os estudos incluídos, foram descritos sistemas que vinculavam o texto produzido às evidências da conversa (STANTON et al., 2025) e frameworks que exigiam verificação de consistência com a fonte antes da integração ao registro (YANG; HA, 2026).

Como orientação prática derivada dos erros descritos, a conferência profissional de rascunhos gerados por IA pode priorizar a fidelidade às fontes originais, o registro de sintomas e riscos, a atribuição correta de falas e comportamentos, a consistência de hipóteses diagnósticas e condutas, a precisão terminológica e a proteção dos dados (SCHWIEGER et al., 2024; OZKARA MENEKSEOGLU et al., 2026; ADHIKARY et al., 2024; YANG; HA, 2026). Essas prioridades não constituem um procedimento de revisão humana validado prospectivamente.

A segurança documental, portanto, não depende exclusivamente do desempenho do modelo, mas também do fluxo sociotécnico em que ele é utilizado. No Smart Notes, consentimento, desidentificação, revisão obrigatória e governança interna integravam a implementação (MCCRUDDEN et al., 2026). Em estudo qualitativo realizado com escriba ambiental em ambiente simulado, confiança, privacidade, consentimento, estigma e particularidades da prática psiquiátrica emergiram como condições importantes para adoção (BOKHARI et al., 2026).

Os vínculos institucionais e financeiros declarados nos estudos de implementação e experiência de uso também devem ser considerados na interpretação. No piloto Anathem, o primeiro autor passou a exercer função na empresa responsável pela ferramenta após o período de coleta do estudo (STANTON et al., 2025). O estudo qualitativo de escriba ambiental recebeu financiamento da Lyrebird Health, tendo os autores declarado que o financiador não participou do desenho, coleta, análise, interpretação ou decisão de publicação (BOKHARI et al., 2026).

No estudo do Smart Notes, todos os autores eram empregados da plataforma avaliada (MCCRUDDEN et al., 2026). Esses vínculos não determinam a validade dos achados, mas constituem informação relevante especialmente na interpretação de medidas autorrelatadas de aceitabilidade, satisfação e qualidade percebida.

5.4 Implicações para pesquisas futuras

Estudos futuros devem priorizar avaliações prospectivas com documentação psiquiátrica real, em diferentes instituições e populações, com critérios explícitos para identificação e classificação de erros clinicamente relevantes e adjudicação independente dos documentos

produzidos. Métricas de similaridade textual, legibilidade ou preferência do avaliador podem caracterizar aspectos documentais, mas não substituem a avaliação direta de fidelidade e segurança clínica.

A supervisão humana também deve ser estudada como intervenção mensurável. São necessários desenhos que comparem diferentes estratégias de revisão e avaliem tempo despendido, tipos de erros identificados, proporção de erros corrigidos e falhas que permanecem após a aprovação profissional. Para favorecer reprodutibilidade, os estudos devem registrar modelo e versão, data de acesso, parâmetros de geração, prompt utilizado e número de execuções, além de considerar reavaliações após atualizações relevantes do sistema.

5.5 Limitações desta revisão narrativa

A heterogeneidade dos estudos quanto aos delineamentos, contextos clínicos, modelos de IA, tarefas documentais e métodos de avaliação impediu a agregação quantitativa dos resultados e a estimativa de uma medida única de acurácia ou erro. Parte das evidências também foi obtida em cenários simulados, materiais sintéticos ou corpus apenas parcialmente psiquiátricos, limitando sua generalização para a documentação clínica real.

O levantamento inicial concentrou-se em PubMed/MEDLINE, LILACS e SciELO, complementadas por ferramentas de descoberta semântica, sem incluir outras bases bibliográficas multidisciplinares, o que pode ter limitado a recuperação de estudos elegíveis. Além disso, a elegibilidade foi restrita a publicações em português, inglês ou espanhol e a textos completos disponíveis gratuitamente. Este último critério pode ter introduzido viés de disponibilidade, uma vez que estudos potencialmente relevantes sem acesso gratuito ao texto completo não puderam ser avaliados integralmente.

A rápida evolução dos modelos de IA também limita a estabilidade temporal dos achados, que devem ser interpretados em relação às versões e condições de uso avaliadas em cada estudo. Além disso, o caráter narrativo da síntese e a ausência de uma classificação formal da qualidade metodológica dos estudos restringem comparações padronizadas entre delineamentos.

Por tratar-se de síntese narrativa baseada em um núcleo delimitado de estudos, a análise não deve ser interpretada como levantamento exaustivo das publicações ou estimativa representativa da frequência de erros em contextos clínicos diversos. As 15 publicações do núcleo analítico foram diferenciadas das referências contextuais, e os limites associados à

seleção por acesso gratuito e à heterogeneidade metodológica foram considerados na interpretação.

6 CONCLUSÃO

A inteligência artificial generativa apresenta potencial para apoiar a documentação em Psiquiatria e saúde mental, mas qualidade formal, legibilidade e fluência não asseguram fidelidade clínica. Foram identificados conteúdos não sustentados, omissões, interpretações inadequadas, erros de atribuição e distorções contextuais em um conjunto heterogêneo de evidências, insuficiente para estimar uma taxa global de acurácia ou de erro. Os resultados favorecem seu uso como apoio à elaboração de rascunhos, com conferência e aprovação profissional antes da incorporação ao prontuário, em vez de utilização documental autônoma. A adoção clínica também demanda governança institucional, proteção dos dados, formação dos profissionais, registro das condições de uso e monitoramento de mudanças de desempenho entre versões. Estudos prospectivos devem avaliar a segurança e a capacidade efetiva das estratégias de supervisão humana para identificar e corrigir erros.

7 INFORMAÇÕES ADICIONAIS

Aspectos éticos. Por utilizar exclusivamente estudos e documentos científicos publicados, sem coleta de dados primários de pessoas ou acesso a informações individuais identificáveis, o trabalho não envolveu pesquisa direta com seres humanos e não foi submetido ao Comitê de Ética em Pesquisa.

Apoio financeiro. Esta revisão não recebeu financiamento específico de agências públicas, comerciais ou sem fins lucrativos.

Conflitos de interesse. Não foram declarados conflitos de interesse.

REFERÊNCIAS

ADHIKARY, P. K. *et al.* Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study. ***JMIR Mental Health***, v. 11, art. e57306, 2024. DOI 10.2196/57306.

ARIHISA, W. *et al.* Comparative analysis of Japanese clinical note styles between physicians and large language models using identical psychiatric cases: quantitative text analysis. ***JMIR Formative Research***, v. 10, art. e85671, 2026. DOI 10.2196/85671.

BALT, E. *et al.* Deductively coding psychosocial autopsy interview data using a few-shot learning large language model. ***Frontiers in Public Health***, v. 13, art. 1512537, 2025. DOI 10.3389/fpubh.2025.1512537.

BLEASE, C.; WORTHEN, A.; TOROUS, J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey. ***Psychiatry Research***, v. 333, art. 115724, 2024. DOI 10.1016/j.psychres.2024.115724.

BOKHARI, S. A. *et al.* Clinician and simulated patient perspectives on ambient AI scribes in psychiatric consultations: a qualitative study. ***Frontiers in Psychiatry***, v. 17, art. 1821065, 2026. DOI 10.3389/fpsyt.2026.1821065.

BOWDEN, S. *et al.* Evaluating the presence of empathic communication in ChatGPT-produced clinical notes using established communication frameworks. ***Cureus***, v. 18, n. 1, art. e102750, 2026. DOI 10.7759/cureus.102750.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). ***Diário Oficial da União***, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 15 out. 2026.

BUCKLEY, P.; WANG, Y.; GOPALAN, P. Artificial intelligence scribes in psychiatry. ***Focus***, v. 23, n. 1, p. 44-48, 2025. DOI 10.1176/appi.focus.20240024.

CASTRO, V. M. *et al.* Psychiatric documentation and management in primary care with artificial intelligence scribe use. ***JAMA Psychiatry***, v. 83, n. 3, p. 281-286, 2026. DOI 10.1001/jamapsychiatry.2025.4303.

CONSELHO FEDERAL DE MEDICINA. Resolução CFM nº 2.454, de 11 de fevereiro de 2026. Normatiza o uso da inteligência artificial na medicina. ***Diário Oficial da União***, Brasília, DF, 27 fev. 2026. Disponível em: https://sistemas.cfm.org.br/normas/visualizar/resolucoes/BR/2026/2454. Acesso em: 15 out. 2026.

GARG, M. *et al.* Leveraging reddit data for context-enhanced synthetic health data generation to identify low self esteem. ***Frontiers in Psychiatry***, v. 16, art. 1726100, 2026. DOI 10.3389/fpsyt.2025.1726100.

GODDARD, K.; ROUDSARI, A. V.; WYATT, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. ***Journal of the American Medical Informatics Association***, v. 19, n. 1, p. 121-127, 2012. DOI 10.1136/amiajnl-2011-000089.

GOLDIE, J. *et al.* Practitioner perspectives on the uses of generative AI chatbots in mental health care: mixed methods study. ***JMIR Human Factors***, v. 12, art. e71065, 2025. DOI 10.2196/71065.

JOHNSON, B. *et al.* General-purpose artificial intelligence use in routine mental health practice among Australian clinicians: mixed methods study. ***Journal of Medical Internet Research***, v. 28, art. e101629, 2026. DOI 10.2196/101629.

KHARKO, A. *et al.* Generative artificial intelligence writing open notes: a mixed methods assessment of the functionality of GPT 3.5 and GPT 4.0. ***Digital Health***, v. 10, art. 20552076241291384, 2024. DOI 10.1177/20552076241291384.

LENZ, E. *et al.* Benchmarking large language models against practicing clinicians on psychopathological assessment. ***npj Digital Medicine***, v. 9, art. 518, 2026. DOI 10.1038/s41746-026-02852-7.

LUO, J. *et al.* Optimization of university counseling consent forms with large language models: multidimensional comparative evaluation. ***Journal of Medical Internet Research***, v. 28, art. e86502, 2026. DOI 10.2196/86502.

MCCRUDDEN, K. E. *et al.* AI-powered documentation for mental health providers: retrospective observational mixed methods study. ***JMIR Formative Research***, v. 10, art. e84628, 2026. DOI 10.2196/84628.

OZKARA MENEKSEOGLU, P. *et al.* Errors in AI-transformed patient-centered mental health documentation written by psychiatrists: qualitative pre-post study. ***JMIR Mental Health***, v. 13, art. e78351, 2026. DOI 10.2196/78351.

RICKMAN, S. Evaluating gender bias in large language models in long-term care. ***BMC Medical Informatics and Decision Making***, v. 25, art. 274, 2025. DOI 10.1186/s12911-025-03118-0.

SCHWARZ, J. *et al.* Psychiatrists' experiences and opinions of generative AI: an exploratory online mixed methods survey in Germany. ***Digital Health***, v. 12, art. 20552076261462618, 2026. DOI 10.1177/20552076261462618.

SCHWIEGER, A. *et al.* Large language models can support generation of standardized discharge summaries: a retrospective study utilizing ChatGPT-4 and electronic health records. ***International Journal of Medical Informatics***, v. 192, art. 105654, 2024. DOI 10.1016/j.ijmedinf.2024.105654.

STANTON, N. *et al.* Evaluating artificial intelligence ambient voice technology as a documentation assistant in psychiatry: proof-of-concept study. ***BJPsych Bulletin***, p. 1-8, 2025. DOI 10.1192/bjb.2025.10186.

WILLIAMS, C. Y. K. *et al.* Evaluating large language models for drafting emergency department encounter summaries. ***PLOS Digital Health***, v. 4, n. 6, art. e0000899, 2025. DOI 10.1371/journal.pdig.0000899.

YANG, M. A.; HA, K. S. A human-governed clinical informatics framework for safe AI-assisted mental health counseling: secondary framework development and requirement mapping study. ***JMIR Formative Research***, v. 10, art. e103345, 2026. DOI 10.2196/103345.