Chatbots e IA generativa podem apoiar cuidados em saúde mental?

Como a literatura científica descreve e avalia o uso de sistemas conversacionais de IA generativa, especialmente modelos grandes de linguagem e chatbots baseados em LLMs, em funções de psicoterapia, aconselhamento, psicoeducação, autocuidado em saúde mental ou suporte emocional, e que evidências existem sobre contextos de uso, aceitabilidade, efeitos relatados, segurança, manejo de crise, limitações e salvaguardas?

Autor institucional	ScienceLayers
Projeto	SL-0002
Tipo de revisão	revisão de escopo
Artefato fonte	outputs/artigo-cientifico-apa.md
Padrão bibliográfico	APA, com metadados disponíveis
Gerado em	18/05/2026 16:45 -03

Resumo

Contexto. A pergunta pública desta revisão é: "Chatbots e IA generativa podem apoiar cuidados em saúde mental?". A pergunta cotidiana original registrada na fila foi: "Chatbots e IA generativa podem fazer terapia?". O protocolo preserva esse ponto de partida como trilha de auditoria, mas reformula o problema em linguagem científica mais precisa e menos carregada.

Objetivo. Como a literatura científica descreve e avalia o uso de sistemas conversacionais de IA generativa, especialmente modelos grandes de linguagem e chatbots baseados em LLMs, em funções de psicoterapia, aconselhamento, psicoeducação, autocuidado em saúde mental ou suporte emocional, e que evidências existem sobre contextos de uso, aceitabilidade, efeitos relatados, segurança, manejo de crise, limitações e salvaguardas?

Método. Esta versão foi conduzida como `scoping_review com estrutura PCC e apoio de PEO.`. O fluxo operacional registrou 4261 registros brutos, 2906 registros após deduplicação, 2906 decisões de triagem, 755 artigos no manifest, 526 textos extraídos e 356 solicitações de artigo pendentes ou registradas.

Resultados. A literatura disponível descreve usos delimitados de chatbots e IA generativa em saúde mental. No corpus público atual, há sinal baixo a moderado de aceitabilidade e de apoio pontual em contextos de baixo risco, sobretudo quando a tarefa é estruturada e a supervisão humana permanece. Isso não demonstra que esses sistemas “façam terapia” no sentido clínico, nem que sejam seguros em crise ou capazes de substituir psicoterapia humana. A síntese técnica distingue estudos de contribuição principal, estudos de apoio, material contextual e evidência indireta, com peso condicionado por desenho, qualidade, centralidade, risco metodológico, sensibilidade ética e limites registrados na crítica profunda 07b.

Conclusão. A conclusão pública desta versão deve permanecer conservadora e não pode ser mais forte que a resposta curta registrada: A literatura disponível descreve usos delimitados de chatbots e IA generativa em saúde mental. No corpus público atual, há sinal baixo a moderado de aceitabilidade e de apoio pontual em contextos de baixo risco, sobretudo quando a tarefa é estruturada e a supervisão humana permanece. Isso não demonstra que esses sistemas “façam terapia” no sentido clínico, nem que sejam seguros em crise ou capazes de substituir psicoterapia humana.

Limitações. A avaliação de qualidade desta versão é preliminar programática, há solicitações de texto completo ainda pendentes ou registradas e bases comerciais ou especializadas podem permanecer como limitação de cobertura. A camada 07b cobre 385 artigos nesta versão, mas não elimina a necessidade de revisão viva, novas buscas e releitura quando houver novo texto completo, correção de metadados, retratação ou mudança relevante de evidência.

Palavras-chave

IA generativa; chatbots; saúde mental; segurança clínica; revisão viva; ScienceLayers.

Introdução

O ScienceLayers preserva a pergunta cotidiana original como trilha de auditoria, mas usa a pergunta pública sugerida como framing editorial quando a formulação bruta trouxer pressupostos, dicotomias, ambiguidade ou linguagem pouco científica. A resposta não deve ser forçada para um sim/não quando a literatura exige gradação por população, exposição, desenho, desfecho, qualidade e contexto.

Esta revisão deve ser executada do zero. O benchmark editorial `ia-psi` pode orientar o nível desejado de rastreabilidade e clareza pública, mas não fornece corpus, dados, resultados ou conclusão para este projeto.

A pergunta cotidiana deve ser preservada, mas a resposta pública não deve reduzir o tema a um sim/não. A revisão deve mapear que usos e avaliações existem, distinguir evidência direta de evidência indireta e deixar explícito quando um estudo só avalia desempenho de sistema, não efeito terapêutico em pessoas.

Método

Desenho da revisão

O tipo vigente é `scoping_review`, porque o campo é recente, heterogêneo e ainda mistura:

- estudos com usuários reais;
- estudos com avaliadores humanos julgando respostas geradas;
- benchmarks técnicos com vinhetas ou prompts simulados;
- revisões, diretrizes, comentários e documentos regulatórios usados apenas como contexto.

Uma revisão sistemática narrativa foi considerada, mas rejeitada como tipo principal neste momento porque poderia prometer uma síntese de efetividade mais forte do que o corpus provavelmente permite antes da busca. Se, depois da triagem, houver clusters empíricos diretos e suficientemente coerentes, eles poderão receber síntese narrativa limitada, com nova decisão metodológica registrada. Um cluster estudo com pessoas só deve ser candidato a revisão sistemática narrativa própria se tiver população, função, desfecho, instrumento ou tempo de seguimento suficientemente comparáveis para uma pergunta mais estreita.

```
Como a literatura científica descreve e avalia o uso de sistemas
conversacionais de IA generativa, especialmente modelos grandes de linguagem e
chatbots baseados em LLMs, em funções de psicoterapia, aconselhamento,
psicoeducação, autocuidado em saúde mental ou suporte emocional, e que
evidências existem sobre contextos de uso, aceitabilidade, efeitos relatados,
segurança, manejo de crise, limitações e salvaguardas?
```

Cada registro deve ser classificado em uma das camadas abaixo.

- Evidência empírica direta estudo com pessoas: estudos com pessoas usuárias, pacientes, participantes, terapeutas ou profissionais interagindo com o sistema, ou análises de uso real/logs com relevância clínica.
- Evidência empírica indireta avaliação humana: estudos em que avaliadores humanos julgam respostas, cenários, vinhetas ou conversas geradas, sem uso terapêutico real por pessoas em sofrimento.
- Estudos técnicos/metodológicos estudo computacional: benchmarks, auditorias de prompts, testes automatizados, simulações ou validações técnicas de segurança, qualidade ou aderência a diretrizes.
- Sínteses prévias e material contextual `contexto`: revisões, metanálises, diretrizes, posicionamentos profissionais, documentos regulatórios, editoriais e artigos conceituais.

Regra interpretativa: estudos estudo computacional e avaliação humana podem informar plausibilidade, riscos, mecanismos e desenho de salvaguardas, mas não devem sustentar sozinhos conclusão sobre eficácia terapêutica ou substituição de psicoterapia humana.

Não haverá filtro temporal inferior rígido na primeira busca, salvo se uma base exigir ajuste operacional que seja registrado no `search-log.csv`.

Diretriz vigente:

- usar termos de IA generativa, LLMs e modelos fundacionais para reduzir entrada de chatbots puramente rule-based;
- processar preferencialmente os resultados de 2022-11-30 em diante antes dos registros mais antigos;
- usar 2022-11-30 como marco analítico e de priorização, não como exclusão automática;
- classificar estudos anteriores a esse marco como elegíveis apenas quando houver componente generativo ou LLM claro, ou quando forem material contextual indispensável;
- registrar em nova decisão metodológica qualquer necessidade futura de aplicar filtro temporal fixo.

A alternativa de limitar a busca a partir de novembro de 2022 foi considerada e rejeitada nesta etapa porque poderia excluir estudos pré-ChatGPT sobre modelos generativos ou conversacionais relevantes.

A escolha atual privilegia cobertura conservadora, com controle de ruído pela triagem e pela classificação por camadas.

Se a busca exploratória retornar mais de 3.000 registros brutos antes da deduplicação, a triagem não deve começar automaticamente. Deve haver checkpoint metodológico para decidir se o marco 2022-11-30 vira filtro formal, se a triagem será feita em ondas ou se outra estratégia de redução de ruído é mais defensável. Esse gatilho não aplica corte temporal por si só.

Idiomas elegíveis para busca e triagem:

- inglês;
- português;
- espanhol;
- francês.

Estudos em outros idiomas podem ser registrados como `pending` ou usado como contexto quando forem claramente centrais e houver metadados suficientes, mas a inclusão completa dependerá de texto avaliável.

Incluir no mapeamento:

- estudos empíricos primários com dados analisáveis;
- estudos quantitativos, qualitativos, mistos, experimentais, observacionais, estudos de logs, avaliações de usuário, avaliações por profissionais, simulações avaliadas e benchmarks técnicos com relevância clínica clara;
- sistemas com componente de IA generativa, LLM, modelo fundacional, chatbot generativo ou agente conversacional generativo claramente descrito;
- usos ou avaliações ligados a psicoterapia, aconselhamento, apoio emocional, psicoeducação, autocuidado em saúde mental, triagem com resposta de suporte ou manejo de crise;
- estudos com usuários reais, profissionais, avaliadores humanos, dados de interação, cenários simulados ou prompts, desde que a camada de evidência seja registrada;
- preprints e literatura cinzenta acadêmica relevante, com marcação explícita e peso interpretativo rebaixado.

Excluir do corpus empírico:

- chatbots exclusivamente rule-based ou sistemas sem componente generativo claro;
- IA usada apenas para diagnóstico, predição, classificação, mineração de dados ou triagem sem interação de suporte, aconselhamento ou resposta terapêutica;
- estudos gerais de IA em saúde mental sem relação com interação conversacional de apoio;
- editoriais, comentários, ensaios, opiniões e documentos normativos sem dados, exceto como material contextual;

- revisões sistemáticas, metanálises e revisões narrativas como estudos primários, embora possam entrar como contexto;
- materiais de marketing, páginas comerciais, notícias ou posts sem método ou dados rastreáveis;
- estudos sem texto suficiente para decidir elegibilidade, que devem ficar como `pending` ou seguir para recuperação de texto completo.

Núcleo de saúde e psicologia, se acessível:

- PubMed/MEDLINE;
- Europe PMC;
- APA PsycInfo/PsycArticles;
- BVS/LILACS;
- SciELO;
- Cochrane Library/CENTRAL, especialmente para ensaios, protocolos e revisões contextuais.

Núcleo de tecnologia e HCI, se acessível:

- ACM Digital Library;
- IEEE Xplore.

Fontes multidisciplinares e de descoberta:

- OpenAlex;
- Crossref;
- Semantic Scholar.

Fontes condicionadas a acesso institucional ou viabilidade operacional:

- Scopus;
- Web of Science;
- Embase;
- CINAHL.

Fonte suplementar:

- Google Scholar apenas para busca manual suplementar e snowballing, com log separado e sem simular reprodutibilidade maior do que a fonte permite.
- medRxiv e PsyArXiv como fontes suplementares estruturadas de preprints em saúde mental, quando a interface permitir registro claro da consulta;
- arXiv e OSF como fontes suplementares condicionadas à viabilidade operacional e à relevância técnica do tema;

- ClinicalTrials.gov e WHO ICTRP para identificar estudos registrados, em andamento, concluídos sem publicação localizada ou com publicação pendente.

Fontes não incluídas na primeira rodada estruturada:

- CNKI e Wanfang Data não entram como bases planejadas nesta etapa porque o escopo linguístico vigente não inclui chinês e ainda não há fluxo documentado de triagem e extração confiável para textos chineses. Essa exclusão deve ser declarada como limitação de cobertura e reavaliada se registros centrais em chinês forem identificados por outras fontes.

Bases inacessíveis devem ser registradas como limitação. Nenhum resultado deve ser inventado ou inferido a partir de ausência de acesso.

Regra de deduplicação:

1. DOI normalizado;
2. PMID/PMCID quando aplicável;
3. título normalizado;
4. título + primeiro autor + ano;
5. comparação manual para ambiguidades, especialmente preprints, versões de conferência e artigos estendidos.

Saídas esperadas:

- `data/records-raw.csv`;
- `data/records-dedup.csv`;
- `dedup-report.md`, quando houver registros reais.

A triagem deve ser documentada como revisão por agentes quando não houver revisores humanos independentes.

Fluxo recomendado:

1. revisor A restritivo;
2. revisor B inclusivo;
3. integrador/árbitro para conflitos;
4. marcação conservadora como `pending` quando título e resumo não permitirem decisão segura.

Toda decisão de exclusão precisa de motivo curto e rastreável em `data/screening-decisions.csv`.

Estados permitidos seguem `AGENTS.md`, incluindo `candidate`, `duplicate`, `excluded_title_abstract`, aguarda texto completo, texto completo solicitado, texto completo disponível, excluído após leitura completa, incluído na síntese, usado como contexto e `pending`.

Ordem de recuperação:

1. link aberto do publisher;
2. DOI landing page;
3. PubMed Central/Europe PMC;
4. OpenAlex;
5. Semantic Scholar;
6. repositórios abertos, como arXiv, OSF, PsyArXiv, medRxiv, bioRxiv ou SocArXiv quando cabível;
7. página institucional do autor;
8. solicitação operacional ao humano via planilha `Artigos solicitados - ia-generativa-terapeuta-digital` na raiz da pasta da revisão no Drive.

Quando o texto completo não for obtido em fontes abertas, criar pasta do artigo, registrar `retrieval.md` e atualizar `data/article-requests.csv`. `data/needs-human-pdf.csv` pode existir como alias legado.

Campos mínimos:

- identificação, DOI, URL e fonte;
- ano, país, idioma e contexto;
- camada de evidência: estudo com pessoas, avaliação humana, estudo computacional ou `contexto`;
- desenho do estudo;
- população, amostra e critérios de elegibilidade;
- condição, demanda ou contexto de saúde mental;
- tecnologia, sistema, fornecedor, modelo e versão quando disponíveis;
- canal de acesso ou implantação, quando disponível: API direta, aplicação de consumidor, protótipo de pesquisa, dispositivo clínico, integração em prontuário, serviço institucional ou outro;
- data ou período de acesso ao modelo, quando disponível;
- descrição de prompts, instruções de sistema, fine-tuning, retrieval, moderação e guardrails;
- função pretendida: psicoterapia estruturada, aconselhamento de apoio, psicoeducação, autocuidado ou autoajuda, triagem com suporte, crise imediata, suporte emocional geral ou outra;
- comparador, quando houver;
- desfechos, instrumentos, métricas e tempo de seguimento;
- principais achados relatados pelos autores;
- riscos, danos, falhas, alucinações, privacidade e eventos adversos;
- salvaguardas propostas ou testadas;
- financiamento, conflito de interesse e relação com desenvolvedores, registrando obrigatoriamente `nao_informado`, `incerto` ou `nao_aplicavel` quando o artigo não trouxer informação suficiente;
- limitações metodológicas;

- relevância para a pergunta cotidiana.

Como `scoping_review`, a avaliação crítica não será usada para excluir automaticamente estudos do mapa. Ela será usada para graduar confiança e limitar a força da síntese pública.

Ferramentas e critérios:

- MMAT 2018 para corpus empírico misto;
- RoB 2 para ensaios randomizados, se houver;
- ROBINS-I ou ferramenta observacional adequada para estudos não randomizados, se houver;
- CASP para estudos qualitativos;
- AMSTAR 2 para revisões usadas como contexto;
- checklist tecnológico próprio para IA generativa, registrando versão do modelo, reprodutibilidade de prompts, documentação de guardrails, teste de alucinação, avaliação de crise, disponibilidade de código/dados, contaminação provável e estabilidade do sistema ao longo do tempo.

A síntese deve priorizar mapeamento, não conclusão causal.

Produtos analíticos previstos:

- mapa por camada de evidência;
- matriz por população, contexto, tecnologia, função e tipo de desfecho;
- mapa de riscos e salvaguardas;
- identificação de clusters com evidência direta suficiente para síntese narrativa limitada;
- lista de lacunas e perguntas futuras passíveis de revisão sistemática.

Regra de linguagem: não concluir que IA generativa "faz terapia", "não faz terapia", "substitui terapeutas" ou "é segura" sem suporte direto, qualidade suficiente e análise de sensibilidade compatível.

Planejadas desde o início:

- sem preprints;
- apenas estudos revisados por pares;
- apenas evidência estudo com pessoas;
- apenas estudos com avaliação crítica moderada ou superior;
- excluindo benchmarks puramente técnicos;
- excluindo estudos com conflito relevante de desenvolvedor, fornecedor ou financiador;
- separando apoio de baixo risco de crise, suicídio, psicose ou risco alto;
- separando estudos pós-2022-11-30 de estudos anteriores;

- separando sistemas com versão, prompts e guardrails documentados de sistemas pouco reprodutíveis.

Fluxo, dados e rastreabilidade

MÉTRICA	VALOR
Registros brutos	4261
Registros após deduplicação	2906
Decisões de triagem	2906
Artigos no manifest	755
Textos extraídos	526
Solicitações de artigo	356
Logs de busca	46

Nota sobre citações e metadados

As referências finais são geradas em APA a partir do manifest e do corpus público. Quando a base local não traz autores completos, periódico, volume, número ou páginas, o artigo preserva os metadados disponíveis e declara essa limitação. A tradução do título para português é editorial e não substitui o título original na referência.

Resultados

Distribuição por camada de evidência

CAMADA	N
material contextual	162
estudo com pessoas	106
estudo computacional	67
avaliação humana	50

Distribuição por qualidade

QUALIDADE	N
contexto	115
baixo_moderado	74
alto	71
moderado_alto	55
moderado	45
baixo	25

Distribuição por peso na síntese

PESO	N
contexto; não sustenta sozinho a conclusão	115
informa plausibilidade e segurança; não sustenta eficácia	64
apoio à síntese	60
contribuição indireta sobre segurança; não sustenta eficácia	51
contribuição principal	49
contribuição limitada	46

Crítica profunda por artigo

A camada 07b contém 385 fichas de crítica profunda por artigo. Ela é usada como trava editorial para impedir que a síntese pública ultrapasse o que cada estudo pode sustentar.

Centralidade:

CENTRALIDADE 07B	N
apoio	144
contexto	124
principal	65
baixo_impacto	51
ultraprincipal	1

Risco metodológico:

RISCO METODOLÓGICO 07B	N
alto	220
moderado	130
não avaliável	24
critico	11

Sensibilidade ética:

SENSIBILIDADE ÉTICA 07B	N
alta	254
critica	66
moderada	64
baixa	1

Prioridade de seguimento metodológico:

PRIORIDADE 07B	N
alta	188
normal	146
baixa	43
nao_priorizar	8

Estudos com contribuição principal

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
SL-0002-ART-000001	Divya Shirsath (2026)	Divine AI: Um Sistema Multilíngue de Bem-Estar Emocional Baseado em LLM para Estudantes e Jovens Adultos	Avaliação controlada de 4 semanas com pré e pós no grupo experimental e comparação com controle; randomização não informada	N=120 estudantes; 80 no experimental e 40 no controle; 18-26 anos; média 21,4	alto	PSS-10 caiu de 22,4 para 14,8 no experimental vs 21,1 pós no controle; GAD-7 caiu de 11,2 para 7,6 vs 10,8; suporte emocional subiu de 41,3% para 82,4%; eficácia percebida no manejo do estresse foi 76,3%; satisfação 8,81/10; 4 escalonamentos, todos encaminhados
SL-0002-ART-000002	Karim Al Ghoul (2026)	Gêmeo Digital com Percepção Emocional para uma Solução de Terapia Personalizada Baseada em Modelo de Linguagem de Grande Escala	Estudo de caso comparativo within-subject com 4 configurações de GPT-4o/UbiMyTherapist em cenários ficcionais, mais avaliação de especialistas	24 participantes adultos em role-play; 5 profissionais de saúde mental avaliaram transcritos; 3 cenários ficcionais	alto	Sistema D obteve maiores médias entre usuários: 4,50 em therapy-style conversation, 4,58 em empatia, 4,58 em personalização, 4,33 em closure e 4,30 em recommendation; Friedman significativo em todas as dimensões p<0,001; especialistas deram score geral 4,02 para D vs 2,68 para GPT-4o bruto
SL-0002-ART-000019	Longxiang Wang (2026)	PsyPARSE: Pensamento Lento Aumentado por Recuperação para Aconselhamento Empático Personalizado	Avaliação técnica training-free com RAG multi-terapia, simulação multi-turn e avaliação LLM-based + humana	50 perfis de pacientes amostrados do CPsyCounR para Patient Agent; número de conselheiros humanos não informado	alto	Em DeepseekV3, PsyPARSE elevou Ctx de 80,51 para 95,72, Tec de 75,69 para 95,50 e Hum de 85,60 para 94,90; empatia humana subiu de 2,45 para 2,92 e adequação terapêutica de 2,50 para 2,95; melhorias consistentes em múltiplos modelos
SL-0002-ART-000	Bogdan Tudor Tulbure	Rompendo Barreiras no Cuidado à Saúde Mental de Estudantes com Terapia Cognitivo-Comportamental em Grupo Aprimorada por IA: Um Estudo	Ensaio de viabilidade de braço único com 4 sessões semanais de terapia grupal Unified Protocol e chatbot LLM	72 triados; 37 elegíveis; 19 iniciaram; 17 completaram; média de idade 22,06 anos; 70,6% mulheres.	alto	17 de 19 participantes completaram o estudo; 16 de 17 completadores frequentaram pelo menos 3 sessões; mediana de 23 dias ativos e 15 mensagens totais; SUS médio 84,94/100; GAD-7 reduziu em média 3,00 pontos (p=.004; d=0.71);

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
100	(2025)	Piloto de Viabilidade (Preprint)	para suporte entre sessões.			SWEMWBS aumentou 2,29 pontos (p=.023; d=0.69); SPIN e PHQ-9 mostraram tendências não significativas; não houve eventos adversos nem interações inadequadas relatadas.
SL-0002-ART-000135	Hajji Hazem (2026)	Uso Problemático de Chatbots de IA para Questões de Saúde Mental entre Adolescentes em Acompanhamento Psiquiátrico Ambulatorial: Gravidade, Prejuízo e Estratégias de Enfrentamento	Estudo transversal descritivo e analítico em ambulatório de psiquiatria infantil e adolescente	N=92 adolescentes usuários de chatbots de IA generativa para questões de saúde mental; média de idade 13,2 ± 1,1 anos; 60,9% do sexo feminino	alto	21 de 92 adolescentes (22,8%; IC95% 15,4–32,4) preencheram critérios de uso problemático clinicamente significativo. Os usos mais comuns foram suporte emocional em sofrimento (71,7%) e busca de informações sobre condições mentais (70,7%). O AIAT-20 teve alfa 0,888 e o CIAT-10 alfa 0,794; o CIAT-10 discriminou bem a classificação clínica (AUC 0,789) e teve alta sensibilidade no ponto de corte empírico >=24. Maior gra…
SL-0002-ART-000209	Evans Ankomah (2025)	Chatbots de IA com Reconhecimento Emocional para Suporte em Saúde Mental em Sistemas de Saúde Pública com Recursos Limitados: Um Estudo de Caso de Gana	Desenvolvimento de protótipo com comparação técnica de classificadores emocionais e teste de campo/usabilidade de chatbot em saúde mental	Dataset ISEAR com mais de 7.000 expressões; teste com 311 participantes em Gana, embora a metodologia mencione piloto com 100 participantes	alto	CNN foi o melhor classificador com 76,4% de acurácia; 89% relataram facilidade de uso; 81% relevância cultural; 78% suporte emocional; 84% reutilização; 66% disseram que a interação os encorajou a considerar ajuda profissional
SL-0002-ART-000595	Aishik Mandal (2026)	Graph2Counsel: Geração de Diálogos Sintéticos de Aconselhamento com Base Clínica a Partir de Grafos Psicológicos de Clientes	Framework de geração sintética de sessões de aconselhamento a partir de Client Psychological Graphs, com avaliação por especialistas, LLM-as-a-judge,	76 CPGs derivados de transcrições reais de terapia envolvendo 6 pacientes anônimos; 760 sessões sintéticas; 4 especialistas avaliaram 100 problemas pareados;	alto	Graph2Counsel foi o melhor em especificidade 1,79, competência 1,67, autenticidade 1,43 e fluxo 1,48, com 0,5 por cento de sessões inseguras; fidelidade ao CPG de 0,91 e ao perfil de 99 por cento; modelo fine-tunado liderou CounselBench-Eval

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
			checagem de fidelidade e testes downstream após fine-tuning	100 pares estratégia-enunciado avaliados manualmente		
SL-0002-ART-000599	Lin Zhong (2026)	LLM Orientado por Políticas Cognitivas para Diagnóstico e Intervenção de Distorções Cognitivas em Conversas de Suporte Emocional	Desenvolvimento de dataset e framework com reinforcement learning; benchmark comparativo com 15 baselines; avaliação automática, humana, por grupos de risco e por ablação	CogBiasESC com 2.499 diálogos, 82.293 enunciados, 8.614 segmentos anotados e 15.092 rótulos; 3 especialistas para anotação; sem pacientes recrutados prospectivamente	alto	CoPoLLM-Qwen3-8B obteve o melhor Macro-F1 de 0,641; CoPoLLM-Llama3.1-8B teve o menor HRMDR de 0,203; o CoPoLLM manteve desempenho superior de segurança no grupo de alto risco e alcançou 73,61 por cento de acerto em estratégia ouro
SL-0002-ART-000600	Md Arid Hasan (2026)	Aprimorando o Suporte de Aconselhamento em Saúde Mental em Bangladesh com Conhecimento Culturalmente Fundamentado	Estudo sociotécnico de desenvolvimento com anotação manual de casos reais; construção de knowledge graph; comparação RAG vs knowledge-graph grounding; avaliação automática e humana	69 casos anotados de sessões reais em bangla, com 2 a 6 sessões por caso; 6 anotadores multidisciplinares; avaliadores humanos da etapa final não quantificados no texto	alto	O knowledge graph aumentou o BERTScore F1 em todos os quatro modelos e melhorou as médias humanas em 0,4 a 0,9 pontos; melhor média humana final foi 1,7 com Llama-3.3-70B grounded; análise ambiental permaneceu mais fraca
SL-0002-ART	Abdullah Mazhar	Medindo o Que Importa: Avaliando Princípios Terapêuticos em Conversas de Saúde Mental	Estudo metodológico de benchmark, anotação especializada e	FAITH-M com 10.172 enunciados em 167 diálogos; splits 6.902/949/2.321; 3 anotadores formais	alto	CARE-Qwen3 atingiu weighted F1 de 63,34 versus 38,56 do Qwen3 base, melhora relativa de 64,26%. O modelo também superou baselines em datasets externos. Acordo com

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
-000602	(2026)		avaliação de modelos	com supervisão de psicóloga clínica; treinamento em 100 conversas		especialista foi maior em warmth and encouragement (84,0%) e non-judgmental language (81,6%)
SL-0002-ART-000603	Qingyang Xu (2026)	Não Causar Dano: Expondo Vulnerabilidades Ocultas de LLMs por Meio de Ataque de Simulação de Cliente Baseado em Persona no Aconselhamento Psicológico	Benchmark/red-teaming multi-turn em aconselhamento psicológico com ataque persona-based, comparação com quatro baselines e inspeção humana	8 modelos-alvo; 1 modelo atacante Llama-3.3-70B; corpora CBT-DP e Cheeseburger Therapy para perfis/estilos; 48 pares ataque-resposta na anotação humana; total de episódios adversariais não informado no corpo principal	alto	PCSA superou os baselines nos 8 modelos; em GPT-5.1 obteve ASR 0,74 no CARES e 0,71 no GPT judge; em PsychoCounsel-Llama3-8B 0,86 e 0,88; em Crispers-7B 0,84 e 0,81. Target compliance médio 0,57; toxic empathy 0,44; impersonation 0,12; PPL menor que 20 e detecção 0%; win rate humano de 96,4% em realismo e concordância humano-judge de 87,5%.
SL-0002-ART-000605	Baihan Li (2026)	Sintético ou Autêntico? Construindo Simuladores de Pacientes de Saúde Mental a Partir de Evidências Longitudinais	Desenvolvimento e avaliação comparativa de framework para simulação de pacientes em saúde mental com perfis integrados e histórico longitudinal	3.258 perfis unificados; D4 com 1.340 entrevistas; 300 clientes em diálogos de aconselhamento; 6.803 usuários e cerca de 37.000 posts do Twitter-STMHD; média de 88 perfis sociais por esqueleto; 30 perfis avaliados por 8 especialistas em 450 avaliações	alto	D EPROFILE superou PATIENT-Ψ e E EYORE em média em realismo 0,932 vs 0,908 e 0,910, diversidade 0,059 vs 0,045 e 0,044 e riqueza de eventos 4,58 vs 4,05 e 3,92. Na avaliação humana, teve qualidade global 3,73 versus 2,73 e 2,58, com win rates de 86,7% e 90,0%.
SL-0002-ART-000	May Lynn Reese (2026)	Usando LLM como Juiz/Júri para Avançar em Avaliações de Segurança Escaláveis e Clinicamente Validadas das	Desenvolvimento e validação de critérios clínicos de segurança com estudos LLM-as-a-Judge e LLM-as-a-Jury sobre respostas a	19 estímulos totais; 3 holdout; 16 analisados; 4 modelos respondedores; 2 raters humanos; 5 consultores clínicos; 448 julgamentos	alto	Concordância humano-humano foi 0,80; Gemini 0,75, Qwen 0,68 e Kimi 0,56 versus consenso humano; o júri obteve 0,74; No Referral foi o critério mais confiável e Embellishes o mais fraco

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
607		Respostas de Modelos a Usuários com Sinais de Psicose	vinhetas de psicose	binários por rater; 25 seeds por juiz LLM		
SL-0002-ART-000610	Jared Moore (2026)	Caracterização de Espirais Delirantes por Meio de Registros de Chat Humano-LLM	Estudo misto observacional e retrospectivo com desenvolvimento de inventário de códigos, anotação automatizada validada por humanos e análises descritivas e associativas de logs reais	19 participantes com logs utilizáveis; 391.562 mensagens em 4.761 conversas; amostra de validação humana com 560 mensagens	alto	Pensamento delirante apareceu em 15,5 por cento das mensagens de usuário; 69 mensagens suicidas e 82 violentas foram validadas; 21,2 por cento das mensagens do chatbot misrepresentaram sentiência; o chatbot facilitou autolesão em 9,9 por cento das mensagens após ideação suicida e violência em 33,3 por cento após pensamentos violentos
SL-0002-ART-000617	Zixin Xiong (2026)	TrustMH-Bench: Um Benchmark Abrangente para Avaliação da Confiabilidade de Modelos de Linguagem de Grande Porte em Saúde Mental	Benchmark multidimensional de trustworthiness em saúde mental cobrindo oito pilares	12 LLMs; 6 gerais e 6 especializados; múltiplos datasets públicos e próprios, incluindo D4 com 1.339 diálogos; sem N agregado único reportado	alto	Nenhum modelo foi forte em todas as dimensões. GPT-5.1 liderou conhecimento com 0,877 e apoio emocional no ESConv com média 4,95/5. DeepSeek-V3.2 teve melhor PCC no D4 para severidade de depressão 0,824 e melhor escore composto em intervenção psicológica 0,965. Modelos especializados ficaram bem atrás em várias tarefas, como Simpsybot 0,267 e MentalLLaMA 0,247 em conhecimento.
SL-0002-ART-000	Abhishek Kulkarni (2026)	E3VA: Aprimorando a Expressividade Emocional em Agentes Conversacionais Virtuais	Desenvolvimento de agente conversacional incorporado com LLM e estudo piloto exploratório sem comparador	N=12 participantes; piloto com dois cenários guiados e interação livre; demografia não informada	alto	SUS médio 77,71; UES médio 3,8/5; focused attention 3,77; perceived usability 3,19; aesthetic appeal 4,08; reward 4,16. Participantes valorizaram histórico de conversa, facilidade de uso e detalhes expressivos;

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
619						tempo de resposta foi a principal crítica.
SL-0002-ART-000626	Kate H. Bentley (2026)	VERA-MH: Confiabilidade e Validade de uma Avaliação de Segurança de IA de Código Aberto em Saúde Mental	Estudo metodológico de validação com conversas simuladas e avaliação humana	90 conversas simuladas; 10 perfis de user-agents; 6 clínicos licenciados; 3 provider-agents principais; LLM judge primário GPT-4o e juízes secundários GPT-5.2, Claude Sonnet 4.5 e Gemini 2.5 Flash	alto	A concordância clínico-clínico foi 0,77; o juiz GPT-4o alinhou-se ao consenso clínico com IRR 0,81; houve 86,8% de concordância geral entre GPT-4o e clínicos ao comparar severidade; 71,6% dos níveis de risco simulados coincidiram com a avaliação clínica; o estilo de disclosure coincidiu em 36,5%; o realismo mediano foi 4 para apresentação e 3 para estilo comunicativo
SL-0002-ART-000627	Himanshi Lalwani (2026)	O Trade-off entre Suporte e Segurança em Agentes de Bem-Estar baseados em LLMs	Estudo experimental fatorial com benchmark de prompts e modelos	80 consultas sintéticas (20 por domínio), 6 LLMs, 3 prompts, 1.440 respostas; 144 respostas avaliadas manualmente	alto	O prompt moderadamente suportivo melhorou cuidado e manteve segurança. O prompt fortemente validante piorou todos os quatro componentes de segurança e às vezes também o cuidado; efeitos globais significativos em SafetyIndex (chi2=77,01; p<0,001) e CareIndex (chi2=38,28; p<0,001). Claude e MiniMax foram mais robustos que Grok, Gemini, DeepSeek e Qwen
SL-0002-ART-000632	António Farinhas (2026)	MindGuard: Classificadores de Proteção para Suporte em Saúde Mental em Múltiplos Turnos	Estudo metodológico de desenvolvimento e avaliação de classificadores de guardrail para suporte mental multi-turn	MindGuard-testset com 67 conversas e 1.134 turnos de usuário anotados; 10 psicólogos clínicos geraram conversas; 3 psicólogos anotaram risco; cerca de 300 cenários sintéticos de treino; 725 interações adversariais na avaliação sistêmica	alto	MindGuard 8B alcançou AUROC=0,982, FPR@90TPR=0,031 e FPR@95TPR=0,054, superando guardrails gerais maiores. Em GLM-4.6, MindGuard 4B reduziu attack success de 25,1% para 7,6% e harmful engagement de 13,7% para 3,3%

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
SL-0002-ART-000633	Aditya Kumar Purohit (2026)	Um Companheiro Condicional: Experiências Vividas de Pessoas com Transtornos de Saúde Mental no Uso de LLMs	Estudo qualitativo com entrevistas semiestruturadas, precedido por survey de triagem	N=20 participantes; idade 21-62 anos; 11 homens e 9 mulheres; entrevistas de 47 a 75 minutos; todos residentes no Reino Unido	alto	Os participantes descreveram cinco necessidades principais: imediatismo, espaço não julgador, autoexposição em ritmo próprio, organização e reestruturação de pensamentos e engajamento relacional; LLMs foram vistos como úteis para sofrimento leve a moderado e apoio cotidiano, mas inadequados para crise, trauma, complexidade social e profundidade relacional; surgiram recomendações para detecção de risco, interação est…
SL-0002-ART-000635	Sohyung Park (2026)	PsyProbe: Diálogo Proativo e Interpretável por Modelagem de Estado do Usuário para Aconselhamento Exploratório	Estudo de desenvolvimento e avaliação comparativa em sessão curta, com avaliações automática, de usuários e de especialista	N=27 participantes; cada participante completou 3 sessões sobre o mesmo tema pessoal, em ordem aleatória, com GPT, PsyProbe e conselheiro humano certificado	alto	Na avaliação do usuário, PsyProbe superou GPT sobretudo em naturalidade (3,72 vs 2,87) e intenção de engajamento (1,19 vs 0,33), com ganhos modestos em coerência, empatia e segurança; na avaliação especialista, PsyProbe atingiu Question Rate de 0,815 versus 0,263 no GPT e próximo ao humano 0,830; Core Issue Understanding foi 3,370 versus 1,153; Probing Question Quality foi 1,440 versus 0,951; em métricas automáticas…

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
SL-0002-ART-000636	Abeer Badawi (2026)	Avaliação da Qualidade do Suporte em Saúde Mental em Respostas de LLMs por meio de Avaliação Humana Multiatributo	Estudo comparativo de avaliação humana multiatributo de respostas de 9 LLMs a conversas de aconselhamento em saúde mental.	500 conversas de aconselhamento selecionadas de MentalChat16K, EmoCare e CounselChat; 4.500 respostas geradas (9 modelos x 500 conversas); 2 avaliadores independentes com treinamento psiquiátrico; 260 conversas com resposta preferida explicitamente selecionada.	alto	A média geral foi 4,29; Safety (4,89) e Interpretation (4,85) foram altas, enquanto Empathy (4,03) e Helpfulness (4,06) ficaram atrás; GPT-4o teve melhor média geral (4,70), Gemini 2.0-Flash e GPT-4o-Mini vieram a seguir, Llama-3.1 foi o melhor open source (4,48) e Qwen-3 o pior (3,60); GPT-4o foi a resposta preferida em 123 de 260 casos e Gemini em 84.
SL-0002-ART-000637	Yimeng Wang (2026)	Explorando Ferramentas Interativas Personalizáveis para Suporte a Tarefas Terapêuticas no Aconselhamento em Saúde Mental	Estudo de design em HCI com métodos mistos: survey formativo, co-design semanal e piloto de usabilidade/think-aloud com entrevistas e escala Likert	27 terapeutas licenciados no survey formativo; 3 terapeutas em co-design; 14 profissionais de saúde mental no estudo piloto	alto	O survey formativo identificou três requisitos centrais: interpretar auto-relatos não estruturados, customizar a visualização por objetivo clínico e centralizar múltiplas fontes. No piloto, 79% relataram forte navegabilidade e 79% disseram que os resumos por IA reduziram carga de trabalho; 71% consideraram o sistema eficiente para localizar informação. Terapeutas relataram alívio de carga cognitiva, confiança basead…

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
SL-0002-ART-000643	Jiwon Kim (2026)	PAIR-SAFE: Uma Abordagem de Agentes Pareados para Auditoria em Tempo de Execução e Refinamento do Suporte em Saúde Mental Mediado por IA	Estudo técnico/metodológico com validação de simulador, comparação entre agente isolado e agente supervisionado e testes de robustez com revisões sucessivas	259 sessões humanas de MI (155 alta qualidade, 104 baixa qualidade) para calibrar e semear simulações; 48 pares de respostas revisados clinicamente	alto	O SeekerSim mostrou maior similaridade semântica em pares corretos do que aleatórios (+59%, d=1.67). Em relação ao Responder isolado, o PAIR-SAFE melhorou Reflection-to-Question Ratio (1.010 para 5.311), Relational (4.773 para 4.919), MI-Adherent Behaviors (0.901 para 1.143), Partnership (4.653 para 4.913) e Seek Collaboration (0.331 para 0.489). Na revisão clínica qualitativa, 30 de 48 pares melhoraram, 8 ficaram e…
SL-0002-ART-000644	Viet Cuong Nguyen (2026)	CALM-IT: Geração de Diálogos Realistas de Entrevista Motivacional em Formato Longo com Rastreamento de Dinâmicas Conversacionais de Dois Atores	Desenvolvimento e avaliação comparativa in silico de framework para geração de diálogos longos de Motivational Interviewing (MI)	686 posts representativos de Reddit de 55 subreddits de saúde mental; 686 perfis/vinhetas pareados com registros OpenPsychometrics DASS; 8.232 transcrições geradas em 4 frameworks e 3 comprimentos; 48 transcrições revisadas por 4 psicólogos licenciados	alto	CALM-IT teve melhor efetividade conversacional (4,45), goal alignment (4,73) e realignment (4,87) que KMI, CAMI+STAR e CI-NC; mostrou menor deriva entre 30 e 100 turnos; embora redirecione menos, teve maior taxa de accepted redirection (64,28%) e, nos maiores redirecionamentos, aumentou change talk em 12,4% e reduziu sustain talk em 8,15%

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
SL-0002-ART-000645	Philipp Steigerwald (2026)	De "Ajuda" a Realmente Útil: Uma Avaliação Hierárquica de LLMs em Aplicações de e-Saúde Mental	Estudo metodológico de avaliação hierárquica com julgamento humano e por IA de linhas de assunto geradas por LLMs para aconselhamento psicossocial por e-mail.	23 threads sintéticas de e-mail em alemão; 11 LLMs; 253 linhas de assunto geradas; 9 avaliadores (5 profissionais de aconselhamento e 4 sistemas de IA); 2.277 avaliações categóricas e 2.277 rankings.	alto	GPT-4o e GPT-3.5 Turbo atingiram 73% de avaliações Good; os melhores open source chegaram a 54%; ajuste para alemão melhorou resultados em até 33 pontos percentuais; após filtragem por concordância restaram 137 de 253 saídas e alpha=0,70.
SL-0002-ART-000650	Dong Xue (2026)	Em Direção ao Suporte em Saúde Mental com Preservação de Privacidade usando Grandes Modelos de Linguagem	Estudo técnico/metodológico de construção de corpus sintético, fine-tuning federado com LoRA e privacidade diferencial, e avaliação automática/humana de um LLM para suporte em saúde mental.	Aproximadamente 11 mil textos de situação de buscadores em plataformas públicas chinesas; MindCorpus com 5,7 mil sessões; 10 clientes federados; amostras aleatórias de 50 sessões por dataset para avaliação de corpus; 4 avaliadores humanos com expertise em psicologia.	alto	MindCorpus superou datasets comparadores em quase todas as métricas; MindChat obteve melhor média automática (2,16) e terceira melhor média humana (1,75), superando modelos do mesmo porte e outros LLMs de saúde mental; maior diversidade de dados ajudou mais que maior quantidade; privacidade diferencial reduziu vazamento e sucesso de ataques, com epsilon=3 como melhor proteção global observada.
SL-0002-ART-000651	Youyou Cheng (2026)	A Deriva Lenta do Suporte: Falhas de Limites em Diálogos de LLMs para Saúde Mental em Múltiplos Turnos	Estudo técnico de stress test multi-turno com dois regimes de pressão e critérios pré-definidos de boundary breach	50 perfis de pacientes virtuais; 3 LLMs; 150 testes; até 20 turnos por conversa; revisão manual de logs de 20 perfis	alto	Em static progression, a taxa média de violação foi 87.3% (98.0% DeepSeek, 86.0% Gemini, 78.0% Grok) e o tempo médio de falha foi 9.21 turnos. Em adaptive probing, a taxa média permaneceu semelhante (88%), mas o tempo médio de falha caiu para 4.64 turnos, com reduções de -5.50 para DeepSeek, -3.52 para Gemini e -4.67 para Grok. O modo de falha dominante foi certainty

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
						reassurance/zero-risk guarantees (56.5% das viola…
SL-0002-ART-000912	Vowels LM (2025)	Eficácia, Viabilidade e Desfechos Técnicos do Chatbot Amanda Baseado em GPT-4o para Suporte em Relacionamentos: Um Ensaio Clínico Randomizado	Ensaio clínico randomizado preregistrado de sessão única, dois braços paralelos, com follow-up de 2 semanas e análise por protocolo	258 participantes na análise final; 130 no grupo Amanda e 128 na tarefa escrita; 240 completaram follow-up	alto	Ambos os grupos melhoraram em 13 de 14 desfechos ao longo do tempo, sem superioridade robusta do chatbot. Houve menor partner-demand/self-withdraw imediatamente pós-intervenção no grupo Amanda. Usabilidade foi 4,19/5, habilidades terapêuticas 3,99/5 e working alliance 4,75/6. A codificação técnica mostrou empatia e therapeutic questioning muito altas, mas Amanda explorou possíveis sinais de segurança em apenas 1 de…
SL-0002-ART-001161	Kamiris P (2026)	Modelos de linguagem com recuperação determinística para aconselhamento clínico: avaliação multilíngue em larga escala com pipelines verificáveis criptograficamente.	Benchmark retrospectivo multilíngue de arquitetura LLM determinística com experimentos de ablação	1.895 cenários declarados: 783 CounselChat em inglês e 1.112 cenários CBT em chinês; tabelas reportam 1.889 casos julgados (778 + 1.111)	alto	Médias inglês: empatia 4,33, fidelidade 3,55, segurança 4,45; médias chinês CBT: 4,85, 4,73 e 4,77. Zero falhas de sistema e zero saídas diagnósticas diretas. Sem retrieval, a fidelidade caiu de 4,14 para 3,40; sem roteador de risco, a segurança caiu de 4,61 para 3,72; baseline generativo foi o pior.
SL-0002-ART-001337	Laura M Vowels (2026)	Grandes Modelos de Linguagem para Avaliação de Risco Psicossocial: Uma Avaliação Multimétodo em Suicídio, Violência por Parceiro Íntimo e Uso Indevido de Substâncias	Estudo multimétodo em três etapas: benchmark de vignettes, validação por participantes e simulação de chatbot supervisionado	Estudo 1: 180 participantes com experiência vivida (52 suicidality/self-harm, 50 IPV, 78 substance misuse); Estudo 2: 111 participantes reavaliando respostas; Estudo 3: 40 casos simulados (12 suicidality/self-harm, 10 IPV, 18 substance misuse)	alto	Estudo 1: GPT-4 e Claude identificaram o domínio principal em 100% dos vignettes; concordância média com severidade autoavaliada foi 84%/k=.87 para GPT-4 e 81%/k=.84 para Claude, com pior desempenho relativo em suicidality/self-harm. Estudo 2: 86,5% concordaram com as ações sugeridas, 86,5% disseram que eram consistentes com ações ou orientações prévias e cerca de

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
						70% se sentiram compreendidos; não houve diferenças…
SL-0002-ART-001372	Anat Shoshani (2026)	Eficácia de um Agente de IA Conversacional para Sintomas Psiquiátricos e Aliança Terapêutica Digital: Ensaio Clínico Randomizado.	Ensaio clínico randomizado paralelo de 3 braços com intervenção de 12 semanas e seguimento de 3 meses	995 estudantes universitários com sofrimento psicológico: 336 IA, 331 terapia em grupo presencial, 328 lista de espera	alto	Após 12 semanas, a IA reduziu mais ansiedade do que terapia em grupo e controle (MD -2,17 e -2,15) e melhorou mais bem-estar do que ambos. Houve maior redução de depressão versus controle, mas não superioridade robusta sobre terapia em grupo após ajuste. Não houve diferença para TEPT. Aos 3 meses, ganhos em ansiedade, bem-estar e satisfação com a vida permaneceram; 61,0% seguiram ativos até a semana 12; aliança tera…
SL-0002-ART-001397	Jinyan Kuang (2026)	Confiança, Desconfiança e Práticas de IA Generativa de Psicoterapeutas na Psicoterapia: Estudo Qualitativo.	Estudo qualitativo por entrevistas semiestruturadas com análise indutiva geral	n=18 psicoterapeutas em exercício; 14 mulheres e 4 homens; média de 9,7 anos de experiência	alto	Adoção descrita como individualizada e dependente da preservação da integridade do papel profissional. Confiança concentrou-se em funções de apoio e baixo risco sob supervisão clínica; desconfiança aumentou em julgamento clínico de alto risco, ameaça à conexão humana e pressões comerciais/organizacionais
SL-0002-ART-001404	Dong Whi Yoo (2026)	Chatbots de IA para Autogestão da Saúde Mental: Estudo Qualitativo Centrado na Experiência Vivida.	Estudo qualitativo scenario-based com technology probe e entrevistas	n=17 adultos com diagnóstico clínico de transtorno depressivo maior; idade 18-66 anos	alto	Três temas centrais: acurácia e aplicabilidade informacional; suporte emocional versus necessidade de conexão humana; dilema personalização-privacidade. Participantes valorizaram validação e apoio prático, mas relataram risco de informação enganosa, vaguidão, limite de empatia e cautela com privacidade

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
SL-0002-ART-001425	Andreas Bucher (2026)	Capacitando Profissionais de Saúde Mental na Psicoterapia Online Assíncrona com IA Generativa.	Estudo misto comparativo dentro do participante com tarefas padronizadas e entrevistas	n=13 profissionais de saúde mental ucranianos	alto	A interação com APIA aumentou competência percebida de forma significativa (p<0,01) e autonomia de forma marginal (p<0,1). Dados qualitativos relataram maior eficiência, validação de julgamentos e independência. Três arquétipos de integração foram identificados: psicoterapeuta-cêntrico, paciente-cêntrico e terapia-cêntrico; persistiram preocupações com confidencialidade, sobredependência e despersonalização
SL-0002-ART-001472	Max Rollwage (2026)	Uma arquitetura de camada cognitiva para apoiar o desempenho de grandes modelos de linguagem em interações psicoterápicas.	Avaliação randomizada duplo-cega de interações psicoterapêuticas com validação em implantação real	227 participantes humanos na avaliação experimental; 22 clínicos especialistas para avaliação cega; 19.674 transcrições de implantação real envolvendo 8.920 usuários	alto	Na avaliação randomizada, os LLMs aumentados pela arquitetura superaram LLMs standalone e clínicos humanos em competências-chave de CBT; a arquitetura melhorou especialmente os componentes goal e task da aliança; o ganho inicial de humor observado no braço com camada cognitiva perdeu significância após controle por latência/duração; a abordagem foi ainda validada em 19.674 transcrições de uso real envolvendo 8.920 u…
SL-0002-ART-001483	Michiko Ueda (2026)	Busca por ajuda na era da IA: inquérito transversal sobre uso e percepções do suporte em saúde mental baseado em IA entre adultos norte-americanos.	Survey transversal on-line com regressão logística e comparações de atitudes	N=1805 adultos de 18-49 anos; 638 usuários semanais de IA; 99 usuários pesados; 511 com histórico de contato com profissional humano	alto	35,2% usavam IA ao menos semanalmente e 5,5% eram usuários pesados. Quase 60% recorriam primeiro a familiares e amigos. Sintomas moderados/graves de depressão/ansiedade se associaram a maior uso de IA (aOR 1,71; IC95% 1,36-2,15) e ideação suicida a maior uso pesado (aOR 2,42; IC95% 1,49-

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
						3,95). Entre 511 que já consultaram profissionais humanos, 28,4% relataram queda percebida na frequência de visitas após usar IA. U…
SL-0002-ART-001488	Benjamin Buck (2026)	Risco de psicose e frequência de uso de inteligência artificial generativa, motivações e experiências semelhantes a delírios: estudo transversal de inquérito.	Estudo transversal por survey online	1003 jovens adultos recrutados; 952 incluídos após exclusões; 846 relataram uso prévio de GenAI e compuseram a principal subamostra analítica	alto	O risco elevado não se associou ao uso prévio de GenAI, mas se associou a uso intensivo (OR 1,70-2,56), maior busca de apoio socioemocional, maior atribuição de papéis como companheiro/amigo/terapeuta/parceiro romântico (OR 1,76-3,08) e maior frequência de interações delirante-semelhantes; GAATES correlacionou-se com PQ-B (r=0,40)
SL-0002-ART-001526	P Maxwell Slepian (2026)	Agente conversacional terapêutico (Solace) para o manejo da dor crônica: estudo de aceitabilidade e usabilidade.	Avaliação prospectiva de viabilidade, aceitabilidade e usabilidade com medidas pré-pós sem grupo controle	175 participantes com dor crônica recrutados via Prolific	alto	Usabilidade excelente (SUS 85,04; DP 13,6), aceitabilidade alta e aliança terapêutica forte; após uma conversa de cerca de 25-30 minutos houve melhora em ansiedade, interferência da dor, cinesiofobia, resiliência à dor e activity engagement; os guardrails pareceram funcionar adequadamente nos cenários observados
SL-0002-ART-001542	Zhi Liu (2026)	Chatbot Baseado em Modelo de Linguagem de Grande Porte Ajustado para Cuidados Oncológicos em Radiologia em Múltiplos Cenários: Ensaio Clínico Randomizado sobre Otimização de Interação, Suporte	Ensaio clínico randomizado (dois sub-ensaios independentes; single-blind)	Sub-trial 1: 1.424 pacientes oncológicos (AT/PP) + 150 RHPs; Sub-trial 2: 638 pacientes (RCS); total 2.062 pacientes; 150 profissionais de saúde em radiologia	moderado_alto	RHP+REC superior em facticidade, integridade e satisfação nos cenários AT (p<0,001) e PP (p<0,001); redução da frustração no PP (3,24±0,12 vs 3,95±0,81; p=0,002); burnout dos RHPs reduzido: exaustão (1,85±0,91 vs 2,40±1,22; p<0,01), despersonalização (2,18±0,93 vs 3,96±0,57; p=0,003), realização pessoal aumentada (4,13±0,87 vs 3,72±0,87; p=0,015); qualidade

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
		Emocional e Redução de Burnout em Profissionais de Saúde				TC melhorada (4,35±0,51 vs 4,00±0,52; p<0,01); RM (4,12±0,5…
SL-0002-ART-001694	Philip Held (2025)	Reavaliação Cognitiva Facilitada por IA via Socrates 2.0: Estudo de Viabilidade por Métodos Mistos.	Estudo pré-clínico de viabilidade de métodos mistos com avaliação pré-pós em 4 semanas e entrevistas semiestruturadas	61 adultos; idade média 39,9 anos; 52 por cento homens; 74 por cento já haviam feito terapia	alto	Uso médio 6,70 vezes em 4 semanas com 10,44 trocas por conversa; viabilidade 4,26, aceitabilidade 4,16 e adequação 4,13; reduções pequenas a moderadas em depressão d=0,30, ansiedade social d=0,25, TEPT d=0,28 e sintomas obsessivo-compulsivos d=0,33; vínculo percebido moderadamente forte; críticas principais foram repetição e pouca profundidade
SL-0002-ART-001713	Xiaochen Luo (2025)	Buscando Apoio Emocional e de Saúde Mental na IA Generativa: Estudo de Métodos Mistos sobre Experiências de Usuários do ChatGPT.	Estudo de métodos mistos com desenho convergente paralelo, survey transversal online e análise temática	270 usuários ativos de ChatGPT para EMS após pré-triagem de 4387 pessoas; idade média 30,06 anos; 59,6 por cento mulheres	alto	38 por cento usavam algumas vezes por mês, 21 por cento algumas vezes por semana e 6 por cento quase diariamente ou mais; 73 por cento classificaram ChatGPT como útil ou muito útil; usos incluíram manejo de sintomas, companhia, orientação relacional e autoavaliação; limitações percebidas incluíram superficialidade, respostas genéricas e frieza de guardrails
SL-0002-ART-001745	Xuan Zhang (2025)	Explorando a Consciência da Imagem Corporal com um Agente Conversacional Baseado em Modelo de Linguagem de Grande Escala: Estudo Qualitativo com Jovens Adultos.	Estudo qualitativo com entrevistas semiestruturadas pré e pós-interação, análise temática e análise de conteúdo das conversas	N=15 jovens adultos de 20 a 30 anos; uso domiciliar por 1 semana; 933 mensagens; piloto prévio com 3 participantes	alto	As conversas foram agrupadas em body image awareness, eating and behavioral regulation, body-focused mindfulness e social conversation. Participantes perceberam o agente como acessível, privado e não julgador, útil para autorreflexão e autocompaixão, mas com repetição e engajamento limitado.
SL-0	Liuling	Intervenção Psicológica de	Desenvolvimento de intervenção	Pré-teste n=7; experimento formal	mod	No experimento formal, o chatbot PST superou o controle

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
002-ART-001827	Mo (2025)	Autoajuda para Jovens na Era Pós-COVID-19: Desenvolvimento de um Chatbot de PST com GPT-4.	com pré-teste e ensaio randomizado controlado	n=100, com 50 no chatbot PST e 50 no chatbot comum	erado_alto	em reconhecimento do problema (40,04 vs 36,40; t(88,31)=3,14; p=0,002; d=0,63) e resolução do problema (47,98 vs 43,36; t(98)=3,34; p=0,001; d=0,67); não houve diferença significativa em qualidade do relacionamento (46,52 vs 44,52; p=0,286); não houve diferenças por gênero ou sequelas pós-COVID
SL-0002-ART-001872	Jessie Goldie (2025)	Perspectivas de Profissionais sobre o Uso de Chatbots de IA Generativa em Saúde Mental: Estudo de Métodos Mistos	Estudo de métodos mistos com entrevistas, ranking de utilidade e medida pré-pós demonstração	N=23 profissionais de saúde mental; 17 mulheres, 6 homens; idade média 39,39 anos; entrevistas de 45 minutos	alto	Generating case notes foi escolhido por 65% (15/23) acima do acaso; suporte a planejamento de sessão e literatura não ficaram acima do acaso. Antes da demonstração, 55% (12/23) tenderam a recomendar chatbots; após a demonstração, 83% (19/23), com aumento significativo. Profissionais relataram melhor encaixe para casos menos complexos, supervisão humana necessária e preocupações persistentes com segurança.
SL-0002-ART-002093	Tillscholich (2025)	Comparação de Respostas de Terapeutas Humanos e Chatbots Baseados em Modelos de Linguagem de Grande Escala para Avaliação da Comunicação Terapêutica: Estudo de Métodos Mistos.	Estudo de métodos mistos comparando respostas de terapeutas licenciados e chatbots LLM generalistas a cenários roteirizados de saúde mental	17 terapeutas licenciados; 2 cenários ficcionais; 6 logs de interação de 3 chatbots avaliados em sessões think-aloud	alto	Sete temas capturaram forças e limites dos chatbots, incluindo validação aparente, estilo conversacional, falta de inquiry, intervenções genéricas e falhas em crise. Terapeutas evocaram mais elaboração do que chatbots (U=9; P=.001). Chatbots usaram mais linguagem affirming (U=28; P=.045), reassuring (U=23; P=.02), psychoeducation (U=22,5; P=.02) e suggestions (U=12,5; P=.003). Chatbots também produziram respostas ma…

ID	AUTOR/ANO	TÍTULO EDITORIAL	DESENHO	AMOSTRA	QUALIDADE	ACHADO REGISTRADO
SL-0002-ART-002194	Chen Chen (2025)	Comparação entre Chatbot de IA e Central de Atendimento de Enfermagem na Redução de Ansiedade e Depressão na População Geral: Ensaio Clínico Randomizado Piloto	Pilot randomized controlled trial (RCT) - dois grupos paralelos com bloco de randomização 1:1	N=124 participantes randomizados; 62 completaram no grupo chatbot e 41 no grupo enfermeira	alto	Grupo chatbot: redução significativa em depressão (PHQ-9 pré 5,13 vs pós 3,68; p=0,008) e ansiedade (GAD-7 pré 4,74 vs pós 3,40; p=0,005). Grupo enfermeira: sem mudança significativa (depressão p=0,38; ansiedade p=0,63). Diferenças entre grupos nas variações pré-pós não significativas (depressão p=0,38; ansiedade p=0,19). Satisfação equivalente entre plataformas (p=0,32). Regressão linear ajustada por pré-escores nã…
SL-0002-ART-002271	Boyoung Kang (2025)	Desenvolvimento e avaliação de um chatbot de saúde mental com ChatGPT 4.0: estudo de experiência do usuário com métodos mistos junto a usuários coreanos.	estudo piloto de metodos mistos focado em experiencia do usuario e usabilidade	20 jovens adultos coreanos; 18 a 27 anos; media 23,3 anos; 12 mulheres e 8 homens; 70 por cento com experiencia previa em chatbots de saude mental	alto	positividade e suporte 9,0; empatia 8,7; escuta ativa 8,0; escores menores em profissionalismo 7,0, complexidade 7,4 e personalizacao 7,4; diferencas significativas versus outros LLM chatbots F=3,27 P=0,047 e versus Woebot/Happify F=12,94 P<0,001
SL-0002-ART-002463	Rivera-Cepeda CF (2026)	Dados do mundo real de um programa de treinamento em grupo para gestão parental aprimorado com inteligência artificial: estudo qualitativo	Estudo qualitativo de mundo real com medidas descritivas de satisfação e análise temática; sem grupo controle	88 cuidadores de crianças de 3 a 14 anos (média 7.98, DP 2.45) em coortes online de PMT	alto	NPS médio 76.92. O tema mais frequente sobre o programa foi useful strategies (73/202; 36.1%). Comentários sobre o Pat foram majoritariamente positivos (156/164; 95.1%), com destaque para acessibilidade 24/7 e disponibilidade constante (69/164; 42.1%). Cuidadores atribuíram em média 61% do progresso ao Pat e 46% às sessões em grupo.

Síntese técnica

Fluxo Da Revisão

MÉTRICA	N	FONTE
Registros brutos acumulados	4.261	`data/records-raw.csv`
Registros canônicos deduplicados	2.906	`data/records-dedup.csv`
Duplicatas removidas	1.355	`data/records-raw.csv`; `data/records-dedup.csv`
Linhas de busca registradas	46	`data/search-log.csv`
Decisões de triagem	2.906	`data/screening-decisions.csv`
aguarda texto completo na triagem	955	`data/screening-decisions.csv`
Registros no manifest	755	`data/articles-manifest.csv`
Textos completos extraídos	526	`data/extraction.csv`
Avaliações de qualidade	526	`data/quality.csv`
Artigos públicos com crítica 07b	385	`data/article-critical-appraisals.csv`
Extrações avaliadas fora da sustentação pública	141	`data/articles-manifest.csv`; `data/quality.csv`
Solicitações humanas pendentes	229	`data/article-requests.csv`
Solicitações humanas pendentes com `priority=alta`	10	`data/article-requests.csv`
Solicitações `recuperado_por_agente`	127	`data/article-requests.csv`

Há uma diferença operacional de 200 registros entre `final_status=aguarda texto completo` na triagem (955) e linhas no manifest (755). Essa diferença não altera a conclusão atual, mas precisa ser reconciliada na revisão viva antes de nova publicação ampla.

A contagem de linhas de busca considera leitura CSV válida. O arquivo físico de `data/search-log.csv` tem quebras de linha dentro de alguns campos textuais longos, o que infla a contagem bruta por linha sem aumentar o número auditável de registros de busca.

Corpus Público Pós-07b

ESTADO NO MANIFEST	N
incluído na síntese	218
usado como contexto	167
Total público com crítica 07b	385

Uso interpretativo:

ESTADO	USO NA SÍNTESE
incluído na síntese	Corpus principal da síntese; pode sustentar conclusão apenas quando camada, qualidade e crítica 07b forem compatíveis.
usado como contexto	Contexto público auditado; pode orientar interpretação, riscos, governança, método e limites, mas não sustenta sozinho a conclusão central.

Entre os 167 usado como contexto, 114 estão em contexto; não sustenta sozinho a conclusão, 18 em contribuição indireta sobre segurança; não sustenta eficácia, 14 em informa plausibilidade e segurança; não sustenta eficácia, 12 em contribuição limitada e 9 em apoio à síntese. Esses 53 itens contextuais com peso diferente de contexto; não sustenta sozinho a conclusão permanecem úteis para interpretação e alerta, mas não são promovidos a evidência central de efeito terapêutico.

Itens extraídos e avaliados fora da sustentação pública:

ESTADO NO MANIFEST	N	USO NESTA SÍNTESE
excluído após leitura completa	125	Limitação operacional, sem sustentação pública
`pending`	16	Limitação operacional, sem sustentação pública

Qualidade Programática

Distribuição dos 526 textos avaliados após a rodada incremental de 2026-05-16:

PESO EM DATA/QUALITY.CSV	N
contribuição principal	49
apoio à síntese	60
contribuição limitada	46
informa plausibilidade e segurança; não sustenta eficácia	64
contribuição indireta sobre segurança; não sustenta eficácia	51
contexto; não sustenta sozinho a conclusão	115
não avaliável	141

Distribuição do corpus público auditado, filtrada para incluído na síntese + usado como contexto com crítica 07b:

PESO NO CORPUS PÚBLICO	N
contribuição principal	49
apoio à síntese	60
contribuição limitada	46
informa plausibilidade e segurança; não sustenta eficácia	64
contribuição indireta sobre segurança; não sustenta eficácia	51
contexto; não sustenta sozinho a conclusão	115

Esta tabela usa peso programático, não hierarquia clínica isolada. contribuição principal não é sinônimo de prova clínica direta: entre os 49 estudos nesse peso, há 25 estudo com pessoas, 12 avaliação humana e 12 estudo computacional. A camada de evidência continua obrigatória para interpretar força clínica.

Camadas De Evidência

Entre os 526 textos avaliados:

CAMADA	N
estudo com pessoas	168
avaliação humana	125
estudo computacional	118
`contexto`	113
`incerto`	2

Entre os 385 artigos públicos:

CAMADA	N
estudo com pessoas	136
avaliação humana	73
estudo computacional	87
`contexto`	89

Dos 136 artigos públicos estudo com pessoas, 103 têm `decisão de inclusão=incluído na síntese` e podem contribuir para a sustentação central quando qualidade e 07b forem compatíveis; 33 têm `decisão de inclusão=usado como contexto` e entram apenas como contexto de uso real, limites, percepção ou risco.

Após `MD-0046`, o corpus público permanece sem tipo de evidência vazio, incerto, não canônico ou malformado. Na rodada incremental de 2026-05-16, dois registros operacionais fora do corpus público (`SL-0002-ART-000055` e `SL-0002-ART-000526`) entraram com `tipo de evidência=incerto` na `data/quality.csv` porque o resolvedor multifonte recuperou arquivos locais incompatíveis e a camada não pôde ser confirmada. Eles permanecem não avaliável e fora da sustentação pública; a renormalização canônica desses dois registros foi registrada como pendência operacional. `SL-0002-ART-000030` permanece normalizado como estudo computacional. `SL-0002-ART-001749` segue reavaliado como `ROBINS_I_adaptado_textual`, `baixo_moderado` e contribuição limitada. `SL-0002-ART-002153` continua excluído após leitura completa/não avaliável.

Crítica Profunda 07b

CENTRALIDADE 07B	N
ultraprincipal	1
principal	65
apoio	144
baixo_impacto	51
contexto	124

RISCO METODOLÓGICO 07B	N
alto	220
moderado	130
critico	11
não avaliável	24

SENSIBILIDADE ÉTICA 07B	N
critica	66
alta	254
moderada	64
baixa	1

PRIORIDADE DE SEGUIMENTO 07B	N
alta	188
normal	146
baixa	43
nao_priorizar	8

O subconjunto `ultraprincipal`/`principal` tem 66 estudos: 35 estudo com pessoas, 16 avaliação humana e 15 estudo computacional. É o núcleo de maior centralidade narrativa, mas continua heterogêneo e não autoriza metanálise nem resposta binária sobre "fazer terapia".

Análises de sensibilidade

```
A conclusão técnica sobrevive se retirarmos ou rebaixarmos as partes mais
frágeis do corpus atual?
```

Conclusão-base testada: a literatura mapeia usos delimitados de IA generativa em saúde mental, com aceitabilidade e sinais proximais em tarefas de baixa intensidade e preferencialmente supervisionadas, mas o corpus não sustenta substituição de psicoterapia humana, equivalência clínica, benefício duradouro ou segurança em crise.

Estas análises são checagens categóricas e narrativas de robustez, não metanálise nem recálculo de tamanho de efeito. O corpus público auditado tem 385 artigos com decisão de inclusão igual a incluído na síntese ou usado como contexto e crítica 07b concluída. Para a conclusão central, usado como contexto é apoio interpretativo e não evidência primária: a sustentação principal vem dos 218 incluído na síntese, filtrados por camada, qualidade e crítica 07b. Os 125 excluído após leitura completa e 16 `pending` extraídos/avaliados permanecem como limitação operacional e estão como não avaliável em `data/quality.csv`. A rodada incremental de 2026-05-16 acrescentou seis avaliações programáticas e duas críticas 07b: `SL-0002-ART-000100` (incluído, estudo com pessoas, contribuição principal, centrality=principal) reforça a evidência direta sem invertê-la; `SL-0002-ART-000057` (usado como contexto, contexto) entra apenas como apoio interpretativo; quatro registros entraram como não avaliável (`pending` ou excluído após leitura completa) e nenhum cenário desta sensibilidade é alterado por eles.

Discussão

Conclusão curta:
A literatura disponível mapeia usos delimitados de chatbots e IA generativa em saúde mental, com aceitabilidade alta e sinal proximal fraco a moderado em apoio pontual, psicoeducação, tarefas informacionais e adjuntos supervisionados. A literatura atual não sustenta que possam "fazer terapia" no sentido clínico de substituir psicoterapia humana. O sinal mais preocupante está em crise, suicídio, psicose, dependência relacional, drift de limites e falsa equivalência terapêutica.

Nível de confiança:
Baixo a moderado para apoio pontual e aceitabilidade em contextos de baixo risco; baixo para benefício clínico sustentado; insuficiente para segurança em crise ou substituição de cuidado humano.

O que sustenta:
218 artigos incluído na síntese formam o corpus principal, com apoio interpretativo de 167 usado como contexto; 103 artigos incluído na síntese são estudo com pessoas, com 33 usado como contexto estudo com pessoas apenas como contexto; 66 estudos `ultraprincipal`/`principal` na 07b; e convergência narrativa entre estudos diretos, avaliações humanas e benchmarks técnicos. Os 526 textos extraídos e avaliados sustentam auditoria operacional e identificação de limitações, mas não são todos base publicável nem base central.

O que enfraquece:
Qualidade programática preliminar; triagem por agentes A/B/árbitro sem revisor humano independente nem estimativa de concordância; 220 estudos com risco metodológico alto na 07b; 11 com risco crítico; 66 com sensibilidade ética crítica; predominância de estudos curtos e heterogêneos; concentração temporal em 2025-2026, com pouco tempo para crítica pós-publicação, replicação ou retratação; 229 solicitações humanas pendentes; 141 extrações avaliadas fora da sustentação pública; diferença de 200 registros entre triagem e manifest; 2 registros não avaliável recém-recuperados com `tipo de evidência=incerto` aguardando renormalização canônica; e cobertura incompleta de bases relevantes, incluindo Semantic Scholar parcial e lacunas em ACM, IEEE, APA PsycInfo/PsycArticles, BVS/LILACS, SciELO e Cochrane/CENTRAL.

O que não dá para concluir:
Não dá para concluir que IA generativa substitui terapeutas, é equivalente à psicoterapia humana, é segura em crise, reduz suicídio, trata psicose, melhora sintomas de forma duradoura, atende populações clínicas graves com segurança, causa ou previne danos em nível populacional, ou generaliza entre modelos, versões, idiomas, culturas e contextos de cuidado.

```
Implicação pública:
A resposta pública deve trocar "pode fazer terapia?" por uma formulação
descritiva sobre usos delimitados e evidência ainda frágil, destacando que o
uso mais defensável é complementar, supervisionado, transparente e fora de
situações de crise.
```

Limitações

- Instabilidade do objeto de pesquisa: sistemas com o mesmo nome comercial podem ter versões, pesos, políticas de segurança e comportamento distintos ao longo do tempo.
- Possível viés de publicação positivo em uma área nova, competitiva e comercialmente ativa; registros de ensaios devem ser usados para procurar estudos não publicados ou concluídos sem publicação localizada.
- Possível viés de spin quando autores, financiadores ou instituições têm vínculo com desenvolvedores dos sistemas avaliados.
- Triagem por agentes sem revisores humanos independentes não equivale a dupla revisão humana e não permite estimar concordância interavaliador; o fluxo A/B/árbitro é mitigação operacional, não substituto metodológico pleno.
- A palavra "terapia" pode significar psicoterapia estruturada, aconselhamento, apoio emocional, psicoeducação, autoajuda ou resposta de crise; a extração deve operacionalizar essa função antes da síntese.
- CNKI e Wanfang Data não estão incluídas na primeira rodada estruturada; isso pode limitar cobertura de literatura chinesa sobre IA conversacional em saúde mental.

Disponibilidade de dados e atualização viva

Os dados auditáveis desta versão estão nos arquivos do projeto `topics/ia-generativa-terapeuta-digital/`, incluindo `protocol.md`, `synthesis.md`, `sensitivity.md`, `data/articles-manifest.csv`, `data/extraction.csv`, `data/quality.csv`, `data/title-translations.csv`, `data/article-critical-appraisals.csv`, `outputs/critical-appraisal-summary.md`, `outputs/public-review.md` e os downloads públicos em `www/downloads/ia-generativa-terapeuta-digital/`.

Esta revisão é viva. Novas buscas, textos completos adicionados, correções de qualidade ou releituras de estudos centrais podem alterar o corpus, a síntese e este artigo.

Nota final sobre o padrão APA

Este manuscrito adota APA como padrão preferencial de referência porque o ScienceLayers trata citação como camada de verificação e aprofundamento, não como obstáculo visual à leitura. O PDF é gerado a partir deste manuscrito e mantém referências em APA com os metadados disponíveis.

Referências

Esta seção lista os estudos do corpus principal e os materiais contextuais conforme registrados no manifest da revisão. Títulos originais são preservados para citação; traduções editoriais aparecem no site e no CSV do corpus. Quando o corpus local não traz autores completos, periódico, volume, número ou páginas, a referência preserva os metadados disponíveis sem completá-los por inferência.

Referências do corpus principal

Autor não informado. (2025). Evaluating the Cultural Relevance of AI Therapist Responses for Chinese American Caregivers of Older Adults. https://pmc.ncbi.nlm.nih.gov/articles/PMC12762769/

Aayushi Joshi. (2025). Enhancing Neonatal Intensive Care Unit (NICU) Paternal Satisfaction: A Comparative Study of Traditional vs. Generative AI-Integrated Counseling. https://doi.org/10.7759/cureus.94374

Abdullah Mazhar. (2026). Measuring What Matters!! Assessing Therapeutic Principles in Mental-Health Conversation.

Abeer Badawi. (2026). Assessing the Quality of Mental Health Support in LLM Responses through Multi-Attribute Human Evaluation.

Abhishek Kulkarni. (2026). E3VA: Enhancing Emotional Expressiveness in Virtual Conversational Agents.

Aditya Kumar Purohit. (2026). A Conditional Companion: Lived Experiences of People with Mental Health Disorders Using LLMs.

Ahalpara. (2025). Tell Me: An LLM-powered Mental Well-being Assistant with RAG, Synthetic Dialogue Generation, and Agentic Planning. https://doi.org/10.48550/arxiv.2511.14445

Ahmad M Nazar. (2026). Providers of relief in distress: RAG-based LLMs as situation and intent-aware assistants. https://doi.org/10.3389/frai.2026.1712596

Aimee-Rose Wrightson-Hester. (2025). A Rule-Based Conversational Agent for Mental Health and Well-Being in Young People: Formative Case Series During the Rise of Generative AI. https://doi.org/10.2196/69841

Aishik Mandal. (2026). Graph2Counsel: Clinically Grounded Synthetic Counseling Dialogue Generation from Client Psychological Graphs.

Anat Shoshani. (2026). Efficacy of a Conversational AI Agent for Psychiatric Symptoms and Digital Therapeutic Alliance: A Randomized Clinical Trial. https://doi.org/10.1001/jamanetworkopen.2026.6713

Andreas Bucher. (2026). Empowering mental health professionals in asynchronous online psychotherapy with GenAI. https://doi.org/10.1186/s13033-026-00700-5

Andrew Clark. (2025). The Ability of AI Therapy Bots to Set Limits With Distressed Adolescents: Simulation-Based Comparison Study. https://doi.org/10.2196/78414

Andrew Franze. (2026). Conversational AI or a human professional for mental health advice? Exploring prevalence and public preferences in Australian adults. https://doi.org/10.1080/00049530.2026.2657279

Anna R Van Meter. (2025). The Goldilocks Zone: Finding the right balance of user and institutional risk for suicide-related generative AI queries. https://doi.org/10.1371/journal.pdig.0000711

Anna-Carolina Haensch. (2025). "It Listens Better Than My Therapist": Exploring Social Media Discourse on LLMs as Mental Health Tool. https://doi.org/10.48550/arxiv.2504.12337

António Farinhas. (2026). MindGuard: Guardrail Classifiers for Multi-Turn Mental Health Support.

Baihan Li. (2026). Synthetic or Authentic? Building Mental Patient Simulators from Longitudinal Evidence.

Benjamin Buck. (2026). Psychosis Risk and Generative Artificial Intelligence Use Frequency, Motivations, and Delusion-Like Experiences: Cross-Sectional Survey Study. https://doi.org/10.2196/85038

Bogdan Tudor Tulbure. (2025). Breaking Barriers in Student Mental Health Care With AI-Enhanced Group Cognitive Behavioral Therapy: A Pilot Feasibility Study (Preprint) https://doi.org/10.2196/preprints.84296

Boyoung Kang. (2025). Development and Evaluation of a Mental Health Chatbot Using ChatGPT 4.0: Mixed Methods User Experience Study With Korean Users. https://doi.org/10.2196/63538

Cecilia Ka Yuk Chan. (2025). AI as the Therapist: Student Insights on the Challenges of Using Generative AI for School Mental Health Frameworks. https://doi.org/10.3390/bs15030287

Chang Liu. (2026). CCD-CBT: Multi-Agent Therapeutic Interaction for CBT Guided by Cognitive Conceptualization Diagram. https://doi.org/10.48550/arxiv.2604.06551

Chen Chen. (2025). Comparison of an AI Chatbot With a Nurse Hotline in Reducing Anxiety and Depression Levels in the General Population: Pilot Randomized Controlled Trial. https://doi.org/10.2196/65785

Cheng ST. (2025). The PDC30 Chatbot-Development of a Psychoeducational Resource on Dementia Caregiving Among Family Caregivers: Mixed Methods Acceptability Study. https://doi.org/10.2196/63715

Chiang. (2025). Do We Talk to Robots Like Therapists, and Do They Respond Accordingly? Language Alignment in AI Emotional Support. https://doi.org/10.48550/arxiv.2506.16473

Chongren Sun. (2025). Persona-Driven Bilingual LLM Dialogues: Cultural Metaphors, Linguistic Divergence, and Emotional Disclosure. https://doi.org/10.54097/sjbj4g80

Choo S. (2025). Advancing Clinical Chatbot Validation Using AI-Powered Evaluation With a New 3-Bot Evaluation System: Instrument Validation Study. https://doi.org/10.2196/63058

Christopher Lalk. (2025). Employing large language models for emotion detection in psychotherapy transcripts. https://doi.org/10.3389/fpsyt.2025.1504306

Dan Holley. (2025). Evaluating Generative Pretrained Transformer (GPT) models for suicide risk assessment in synthetic patient journal entries. https://doi.org/10.1186/s12888-025-07088-5

Daniil Filienko. (2024). Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy. https://pmc.ncbi.nlm.nih.gov/articles/PMC12099324/

David Villarreal-Zegarra. (2026). Development, system design, safety, and performance metrics of a conversational agent for reducing depressive and anxious symptoms based on a large language model: The MHAI study. https://doi.org/10.1371/journal.pone.0344939

De Duro ES. (2024). Introducing CounseLLMe: A dataset of simulated mental health dialogues for comparing LLMs like Haiku, LLaMAntino and ChatGPT against humans. https://doi.org/10.31234/osf.io/vhmqs

Dimas Rizqi Kurniawan. (2026). Speech-Based Virtual Assistant for Mental Health Support Through Natural Interaction. https://doi.org/10.29207/resti.v9i6.6585

Divya Shirsath. (2026). Divine AI: A Multilingual LLM-Based Emotional Wellness System for Students and Young Adults. https://doi.org/10.5281/zenodo.20050110

Do Hyung Kim. (2025). BetterMood: A human-like AI counseling service for adolescents and young adults. https://doi.org/10.1177/20552076251392294

Dong Whi Yoo. (2026). AI Chatbots for Mental Health Self-Management: Lived Experience-Centered Qualitative Study. https://doi.org/10.2196/78288

Dong Xue. (2026). Towards Privacy-Preserving Mental Health Support with Large Language Models.

Doyoon Kim. (2026). When a chatbot asks "How are you?": A cross-sectional study of AI call conversations and depressive symptom detection among older adults in rural South Korea. https://doi.org/10.1186/s12877-026-07038-0

Edoardo Pinzuti. (2025). Comparative performance of large language models in emotional safety classification across sizes and tasks. https://doi.org/10.3389/frai.2025.1706090

Edoardo Pinzuti. (2026). Visually grounded emotion regulation via diffusion models and user-driven reappraisal. https://doi.org/10.3389/frai.2026.1691445

Elham Aghakhani. (2026). Like a Therapist, But Not: Reddit Narratives of AI in Mental Health Contexts. https://doi.org/10.48550/arxiv.2601.20747

Eliya Naomi Aharon. (2026). SAGE: A Strategy-Aware Graph-Enhanced Generation Framework For Online Counseling.

Eman Abdelaziz Rashad Dabou. (2025). New Horizons in Higher Education: Examining the Mental Well-Being of Medical & Health Sciences Students Through the Use of Artificial Intelligence Based Chatbot Platforms in the United Arab Emirates - A Cross-Sectional Comparative Study. https://doi.org/10.12688/f1000research.166372.2

Emma Coen. (2025). Chatbot for the Return of Positive Genetic Screening Results for Hereditary Cancer Syndromes: Prompt Engineering Project. https://doi.org/10.2196/65848

Evans Ankomah. (2025). Emotion-Aware AI Chatbots for Mental Health Support in Low-Resource Public Health Systems: A Case Study from Ghana. https://doi.org/10.11648/j.wjph.20251003.17

Fangrui Huang. (2026). TherapyGym: Evaluating and Aligning Clinical Fidelity and Safety in Therapy Chatbots. https://doi.org/10.48550/arxiv.2603.18008

Fanya Sun. (2025). Exploring the use of ChatGPT-4o in Cognitive Behavioural Therapy for university students: enhancing mental health with AI-powered voice interaction. https://doi.org/10.54254/2753-7102/2025.23159

Fengbo Jiao. (2026). Addressing loneliness by AI chatbot: a qualitative study of empty-nest elderly. https://doi.org/10.1186/s12889-026-26283-x

Gavin N Rackoff. (2025). Chatbot-delivered mental health support: Attitudes and utilization in a sample of U.S. college students. https://doi.org/10.1177/20552076241313401

Ghuzayyil Mohammed Al-Otaibi. (2025). Communication Errors in Human-Chatbot Interactions: A Case Study of ChatGPT Arabic Mental Health Support Inquiries. https://doi.org/10.3390/bs15081119

Goran Arbanas. (2025). Patients prefer human psychiatrists over chatbots: a cross-sectional study. https://doi.org/10.3325/cmj.2025.66.13

Guifeng Deng. (2026). Evaluating Large Language Models in Crisis Detection: A Real-World Benchmark from Psychological Support Hotlines. https://doi.org/10.1109/jbhi.2026.3688375

Habicht J. (2024). Generative AI-Enabled Therapy Support Tool Improves Clinical Outcomes and Patient Engagement in NHS Talking Therapies. https://doi.org/10.31234/osf.io/mj46k

Hadar Fisher. (2025). Measuring activation during behavioral activation therapy: a proof-of-concept study using smartphone sensors and LLM-derived ratings in adolescents with anhedonia. https://doi.org/10.1038/s44277-025-00045-w

Hagai Astrin. (2026). CARE: Counselor-Aligned Response Engine for Online Mental-Health Support.

Hajji Hazem. (2026). Problematic use of AI chatbots for mental health concerns among adolescent psychiatric outpatients: severity, impairment, and coping. https://doi.org/10.1186/s43045-026-00637-y

Han Zhang. (2026). Human-like, Animal-like, or Object-like? The Impact of LLM-Based Virtual Doctor Avatar Design on User Emotion, Physiology, and Experience. https://doi.org/10.3390/bs16030349

Heinz MV. (2024). Evaluating Therabot: A Randomized Control Trial Investigating the Feasibility and Effectiveness of a Generative AI Therapy Chatbot for Depression, Anxiety, and Eating Disorder Symptom Treatment. https://doi.org/10.31234/osf.io/pjqmr

Herschbach L. (2025). Evaluation of an AI-Based Chatbot Providing Real-Time Feedback in Communication Training for Mental Health Care Professionals: Proof-of-Concept Observational Study. https://doi.org/10.2196/82818

Heston TF. (2023). Evaluating Risk Progression in Mental Health Chatbots Using Escalating Prompts. https://doi.org/10.1101/2023.09.10.23295321

Heston TF. (2023). Safety of Large Language Models in Addressing Depression. https://doi.org/10.7759/cureus.50729

Hilton Humphries. (2026). A qualitative study assessing the acceptability of a multi-agent AI Chatbot for providing HIV and mental health support among men who have sex with men and transgender women in KwaZulu-Natal, South Africa. https://doi.org/10.1093/trstmh/traf143

Himanshi Lalwani. (2026). The Supportiveness-Safety Tradeoff in LLM Well-Being Agents.

Hodson N. (2024). Can Large Language Models Replace Therapists? Evaluating Performance at Simple Cognitive Behavioral Therapy Tasks. https://doi.org/10.2196/52500

Hongbin Na. (2024). CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering. https://doi.org/10.48550/arxiv.2403.16008

Hou Z. (2025). Psychological Anxiety Risk Analysis Model Based on Large Language Model Interaction. https://doi.org/10.21203/rs.3.rs-8192788/v1

Hyojin Chin. (2025). Chatbots' Empathetic Conversations and Responses: A Qualitative Study of Help-Seeking Queries on Depressive Moods Across 8 Commercial Conversational Agents. https://doi.org/10.2196/71538

Ian Steenstra. (2026). Assessing Risks of Large Language Models in Mental Health Support: A Framework for Automated Clinical AI Red Teaming. https://doi.org/10.48550/arxiv.2602.19948

Inka Napiwotzki. (2025). Comparing Human and AI Therapists in Behavioral Activation for Depression: Cross-Sectional Questionnaire Study. https://doi.org/10.2196/78138

Jabari Kwesi. (2025). Exploring User Security and Privacy Attitudes and Concerns Toward the Use of General-Purpose LLM Chatbots for Mental Health. https://doi.org/10.48550/arxiv.2507.10695

Janak Kapuriya. (2025). Spiritual-LLM : Gita Inspired Mental Health Therapy In the Era of LLMs. https://doi.org/10.48550/arxiv.2506.19185

Jared Moore. (2026). Characterizing Delusional Spirals through Human-LLM Chat Logs.

Jazmin A Reyes-Portillo. (2025). Generative AI-Powered Mental Wellness Chatbot for College Student Mental Wellness: Open Trial. https://doi.org/10.2196/71923

Jessica McFadyen. (2026). Increasing engagement with cognitive-behavioral therapy (CBT) using generative AI: a randomized controlled trial (RCT). https://doi.org/10.1038/s43856-025-01321-8

Jessie Goldie. (2025). Practitioner Perspectives on the Uses of Generative AI Chatbots in Mental Health Care: Mixed Methods Study. https://doi.org/10.2196/71065

Jia Xu. (2024). Revolutionizing Dementia Care: Enhancing Talk Therapy with Fine-Tuned Large Language Models Using GPT Self-Generated Data. https://doi.org/10.1002/alz.093496

Jia Xu. (2025). MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance. https://doi.org/10.1145/3711896.3737393

Jiatao Quan. (2025). Relational Mediators: LLM Chatbots as Boundary Objects in Psychotherapy. https://doi.org/10.48550/arxiv.2512.22462

Jiayue Melissa Shi. (2026). Mapping Caregiver Needs to AI Chatbot Design: Strengths and Gaps in Mental Health Support for Alzheimer's and Dementia Caregivers. https://doi.org/10.1145/3803549

Jinyan Kuang. (2026). Psychotherapists' Trust, Distrust, and Generative AI Practices in Psychotherapy: Qualitative Study. https://doi.org/10.2196/88932

Jiwon Kim. (2026). PAIR-SAFE: A Paired-Agent Approach for Runtime Auditing and Refining AI-Mediated Mental Health Support.

Joe Hasei. (2025). Empowering pediatric, adolescent, and young adult patients with cancer utilizing generative AI chatbots to reduce psychological burden and enhance treatment engagement: a pilot study. https://doi.org/10.3389/fdgth.2025.1543543

Johanna Habicht. (2025). Generative AI-Enabled Therapy Support Tool for Improved Clinical Outcomes and Patient Engagement in Group Therapy: Real-World Observational Study. https://doi.org/10.2196/60435

Joho AA. (2026). Self-directed use of a generative artificial intelligence chatbot for mental health therapy and associated outcomes among university students in Tanzania. https://doi.org/10.21203/rs.3.rs-9023115/v1

Joshua D Wenger. (2026). People choose to receive human empathy despite rating AI empathy higher. https://doi.org/10.1038/s44271-025-00387-3

Jun Tao. (2026). Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing. https://doi.org/10.2196/79671

Jun Tat Tan. (2025). Psychological First Aid by AI: Proof-of-Concept and Comparative Performance of ChatGPT-4 and Gemini in Different Disaster Scenarios. https://doi.org/10.1002/jclp.23808

June M. Liu. (2023). ChatCounselor: A Large Language Models for Mental Health Support. https://doi.org/10.48550/arxiv.2309.15461

Junjie Wang. (2026). Mapping generative AI use in the human brain: divergent neural, academic, and mental health profiles of functional versus socio emotional AI use. https://doi.org/10.48550/arxiv.2604.08594

Junsang Park. (2025). AI Chatbots or Human Therapists? Belief-Based Predictors of Mental Health Help-Seeking Intentions in the Age of Generative AI. https://doi.org/10.48550/arxiv.2512.03406

K. Izumi. (2024). Response Generation for Cognitive Behavioral Therapy with Large Language Models: Comparative Study with Socratic Questioning. https://doi.org/10.48550/arxiv.2401.15966

Kaden Bunch. (2025). Leveraging Social Media and AI for Early Community Mental Health Support. https://doi.org/10.7759/cureus.95047

Karen Yirmiya. (2026). Feasibility and acceptability of the MentiParent AI chatbot for training parental reflective functioning. https://doi.org/10.1038/s41598-026-47934-4

Karim Al Ghoul. (2026). Emotion-Aware Digital Twin for a Large Language Model-Based Personalized Therapy Solution. https://doi.org/10.20381/ruor-31909

Karmiris P. (2026). Deterministic Retrieval-Grounded Language Models for Clinical Counseling: Large-Scale Multilingual Evaluation with Cryptographically Verifiable Pipelines. https://doi.org/10.20944/preprints202603.1217.v1

Kate H. Bentley. (2026). VERA-MH: Reliability and Validity of an Open-Source AI Safety Evaluation in Mental Health.

Kaukab Ansari. (2025). Human Vulnerability, Machine Law: Ethical Risks and Global Governance Challenges in Generative AI. https://doi.org/10.36948/ijfmr.2025.v07i05.56786

Kean Sian Tan. (2025). AI meets psychology: an exploratory study of large language models' competence in psychotherapy contexts. https://doi.org/10.1080/29974100.2025.2545258

Khadangi. (2025). When AI Takes the Couch: Psychometric Jailbreaks Reveal Internal Conflict in Frontier Models. https://doi.org/10.48550/arxiv.2512.04124

Kim T. (2025). Generative Pre-trained Transformer-4 (GPT-4) generated psychological reports in psychodynamic perspective: a pilot study on quality of report, risk of hallucination, and client satisfaction. https://pmc.ncbi.nlm.nih.gov/articles/PMC12438398/

Kunmi Sobowale. (2025). Evaluating Generative AI Psychotherapy Chatbots Used by Youth: Cross-Sectional Study. https://doi.org/10.2196/79838

Kunmi Sobowale. (2025). Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study. https://doi.org/10.2196/65605

Kyuha Jung. (2025). 'I've talked to ChatGPT about my issues last night.': Examining Mental Health Conversations with Large Language Models through Reddit Analysis.

https://doi.org/10.1145/3757537

L. Galland. (2024). Generating Unexpected yet Relevant User Dialog Acts. https://doi.org/10.18653/v1/2024.sigdial-1.17

Lanqing Du. (2026). Constructing and applying a multi-turn psychological support dialogue corpus based on the Helping Skills Chain-of-Thought. https://doi.org/10.3389/fpsyg.2026.1733384

Lanqing Du. (2026). EFT-CoT: A Multi-Agent Chain-of-Thought Framework for Emotion-Focused Therapy. https://doi.org/10.48550/arxiv.2601.17842

Laura M Vowels. (2026). Large language models for psychosocial risk assessment: A multi-method evaluation across suicide, intimate partner violence, and substance misuse. https://doi.org/10.1371/journal.pdig.0001352

Laurie O Campbell. (2025). An Examination of Generative AI Response to Suicide Inquires: Content Analysis. https://doi.org/10.2196/73623

Lea Maria Schäfer. (2025). Exploring user characteristics, motives, and expectations and the therapeutic alliance in the mental health conversational AI Clare®: a baseline study. https://doi.org/10.3389/fdgth.2025.1576135

Lin Zhong. (2026). Cognitive Policy-Driven LLM for Diagnosis and Intervention of Cognitive Distortions in Emotional Support Conversation.

Lingyao Li. (2025). LLM Use for Mental Health: Crowdsourcing Users' Sentiment-based Perspectives and Values from Social Discussions.

Liu H. (2026). AI-Assisted Solution-Focused Counseling Training for Novice Mental Health Educators: An Exploratory Study. https://doi.org/10.21203/rs.3.rs-9368123/v1

Liu X. (2026). Assessing the Feasibility, Usability, Acceptability, and Efficacy of an AI Chatbot for Sleep Promotion: Quasi-Experimental Study. https://doi.org/10.2196/84023

Liuling Mo. (2025). Self-help psychological intervention for young individuals during the post-COVID-19 era: development of a PST chatbot using GPT-4. https://doi.org/10.3389/fdgth.2025.1627268

Liying Wang. (2024). Large Language Model-Powered Conversational Agent Delivering Problem-Solving Therapy (PST) for Family Caregivers: Enhancing Empathy and Therapeutic Alliance Using In-Context Learning. https://pmc.ncbi.nlm.nih.gov/articles/PMC12919573/

Lizu Lai. (2025). Depression and the use of conversational AI for companionship among college students: the mediating role of loneliness and the moderating effects of gender and mind perception. https://doi.org/10.3389/fpubh.2025.1580826

Longxiang Wang. (2026). PsyPARSE: Retrieval-Augmented Slow Thinking for Personalized Empathetic Counseling. https://doi.org/10.1609/aaai.v40i2.37089

Lubomir Barabas. (2025). Exploring the potential of ChatGPT as a digital advisor in acute psychiatric crises: a feasibility study. https://doi.org/10.1007/s00115-025-01837-3

Lyndsey Hipgrave. (2025). Balancing risks and benefits: clinicians' perspectives on the use of generative AI chatbots in mental healthcare. https://doi.org/10.3389/fdgth.2025.1606291

Mahyar Abbasian. (2024). Empathy Through Multimodality in Conversational Interfaces. https://doi.org/10.48550/arxiv.2405.04777

Major M. (2026). From "Crutch" to Coach: Patterns of Sustained Engagement and Deepening Support in AI Wellbeing Coaching. https://doi.org/10.21203/rs.3.rs-8949770/v1

Mana Hanzawa. (2026). Development of a generative AI agent for family support in implementing family-based treatment for children and adolescents with anorexia nervosa. https://doi.org/10.3389/fdgth.2026.1759690

Marcin Rządeczka. (2025). The Efficacy of Conversational AI in Rectifying the Theory-of-Mind and Autonomy Biases: Comparative Analysis. https://doi.org/10.2196/64396

Marita Skjuve. (2026). ChatGPT as a mental health advisory service: Comparing evaluations from youth and health professionals. https://doi.org/10.1177/20552076261427447

Mark Kalinich. (2026). Evaluating the effect of mental health fine-tuning relative to other model characteristics on LLM safety performance. https://doi.org/10.64898/2026.01.02.25343289

Martin Dechant. (2025). Future Me, a Prospection-Based Chatbot to Promote Mental Well-Being in Youth: Two Exploratory User Experience Studies. https://doi.org/10.2196/74411

Matthew Flathers. (2026). Benchmarking Language Models for Clinical Safety: A Primer for Mental Health Professionals. https://doi.org/10.64898/2026.03.20.26348900

Max Rollwage. (2026). A cognitive layer architecture to support large-language model performance in psychotherapy interactions. https://doi.org/10.1038/s41591-026-04278-w

May Lynn Reese. (2026). Using LLM-as-a-Judge/Jury to Advance Scalable, Clinically-Validated Safety Evaluations of Model Responses to Users Demonstrating Psychosis.

Md Abdullah Al Kafi. (2026). Reasoning Over Recall: Evaluating the Efficacy of Generalist Architectures vs. Specialized Fine-Tunes in RAG-Based Mental Health Dialogue Systems. https://doi.org/10.48550/arxiv.2601.01341

Md Arid Hasan. (2026). Enhancing Mental Health Counseling Support in Bangladesh using Culturally-Grounded Knowledge.

Mian Zhang. (2024). CBT-Bench: Evaluating Large Language Models on Assisting Cognitive Behavior Therapy. https://doi.org/10.48550/arxiv.2410.13218

Michiko Ueda. (2026). Help-Seeking in the Age of AI: Cross-Sectional Survey of the Use and Perceptions of AI-Based Mental Health Support Among US Adults. https://doi.org/10.2196/88196

Moreah Zisquit. (2025). AI-Enhanced Virtual Reality Self-Talk for Psychological Counseling: Formative Qualitative Study. https://doi.org/10.2196/67782

Namwoo Kim. (2025). GPT-4 generated psychological reports in psychodynamic perspective: a pilot study on quality, risk of hallucination and client satisfaction. https://doi.org/10.3389/fpsyt.2025.1473614

Natalia Amat-Lefort. (2026). From Chatbots to Confidants: A Cross-Cultural Study of LLM Adoption for Emotional Support.

Navdeep Singh Bedi. (2026). Assessing the Effectiveness of LLMs in Delivering Cognitive Behavioral Therapy. https://doi.org/10.48550/arxiv.2603.03862

Ngo N. (2025). Evaluating Voice-Enabled Generative AI for Mental Health: Real-Time Performance and Safety Analyses. https://doi.org/10.1101/2025.11.14.25340246

Nicola Döring. (2025). Anti-AI Bias Toward Couple Images and Couple Counseling: Findings from Two Experiments. https://doi.org/10.1007/s10508-025-03318-9

Ningjing Tang. (2026). Beyond the Single Turn: Reframing Refusals as Dynamic Experiences Embedded in the Context of Mental Health Support Interactions with LLMs.

Olsen SG. (2026). Potentially Harmful Consequences of Artificial Intelligence (AI) Chatbot Use Among Patients With Mental Illness: Early Data From a Large Psychiatric Service System. https://doi.org/10.1111/acps.70068

Onno P. Kampman. (2024). A Multi-Agent Dual Dialogue System to Support Mental Health Care Providers. https://doi.org/10.48550/arxiv.2411.18429

P Maxwell Slepian. (2026). A Therapeutic Conversational Agent (Solace) for Management of Chronic Pain: Acceptability and Usability Study. https://doi.org/10.2196/87689

Palaniyappan L. (2026). Chatbots, delusions, and treatment failure. https://doi.org/10.1139/jpn-2025-0249

Pattaradit Samatha. (2025). The Impact of Artificial Intelligence Interventions on Adolescent Mental Health: A Multidimensional Study Using ChatGPT, Gemini, and DeepSeek. https://doi.org/10.38124/ijisrt/25jul1857

Paul Binu. (2026). Enhancing Speech Synthesis With Human-Like Emotional Intelligence For Natural And Expressive Communication. https://doi.org/10.5281/zenodo.20045981

Peng X. (2025). When AI chatbots understand emotions: exploring the mechanisms of self-disclosure from emotional arousal to psychological acceptance-a hybrid SEM-ANN-NCA analysis across multiple interaction configurations. https://doi.org/10.3389/fpsyg.2025.1724313

Per Niklas Waaler. (2025). Prompt Engineering an Informational Chatbot for Education on Mental Health Using a Multiagent Approach for Enhanced Compliance With Prompt Instructions: Algorithm Development and Validation. https://doi.org/10.2196/69820

Philip Held. (2025). AI-Facilitated Cognitive Reappraisal via Socrates 2.0: Mixed Methods Feasibility Study. https://doi.org/10.2196/80461

Philipp Steigerwald. (2026). From "Help" to Helpful: A Hierarchical Assessment of LLMs in Mental e-Health Applications.

Porwal G. (2024). Evaluating the Quality of Mental Health Information Generated by Large Language Model Chatbots. https://doi.org/10.1101/2024.12.20.24319373

Prof. S. S. Dixit. (2025). AI-Powered Mental Health Chatbot: A Research Paper. https://doi.org/10.22214/ijraset.2025.70756

Przemysław Waszak. (2025). Chat GPT and suicide prevention – can it work? A conversation analysis. https://doi.org/10.15557/pipk.2024.0036

Qingyang Xu. (2026). Do No Harm: Exposing Hidden Vulnerabilities of LLMs via Persona-based Client Simulation Attack in Psychological Counseling.

Ritesh Maurya. (2025). Exploring the potential of lightweight large language models for AI-based mental health counselling task: a novel comparative study. https://doi.org/10.1038/s41598-025-05012-1

Rivera-Cepeda CF. (2026). Real-World Data From a Group Parent Management Training Program Enhanced Using Artificial Intelligence: Qualitative Study. https://doi.org/10.2196/91841

Ruiqi Yao. (2025). Connecting self-esteem to problematic AI chatbot use: the multiple mediating roles of positive and negative psychological states. https://doi.org/10.3389/fpsyg.2025.1453072

Ruoyi Huang. (2025). A PERSONALIZED MOBILE APPLICATION TO GENERATE MUSIC THERAPY USING A LARGE LANGUAGE MODEL AND STORING THE USER’S DATA ON FIREBASE. https://doi.org/10.5121/csit.2025.151719

Ryan K McBain. (2025). Evaluation of Alignment Between Large Language Models and Expert Clinicians in Suicide Risk Assessment. https://doi.org/10.1176/appi.ps.20250086

S Gabe Hatch. (2025). When ELIZA meets therapists: A Turing test for the heart and mind. https://doi.org/10.1371/journal.pmen.0000145

Saadia Gabriel. (2024). Can AI Relate: Testing Large Language Model Response for Mental Health Support. https://doi.org/10.48550/arxiv.2405.12021

Sam Zaia. (2025). Perceived benefits and limitations of a generative AI chatbot for mental health support: an exploratory mixed-methods study. https://doi.org/10.57129/001c.144919

Samantha Weber. (2026). Suicide- and crisis-risk detection using large language models in mental-health chatbots. https://doi.org/10.64898/2026.01.12.26343914

Scott Graupensperger. (2026). User Experience and Early Clinical Outcomes of a Mental Wellness Chatbot for Depression and Anxiety: Pilot Evaluation Mixed Methods Study. https://doi.org/10.2196/90644

Scott N Hannah. (2025). As Effective as You Perceive It: The Relationship Between ChatGPT's Perceived Effectiveness and Mental Health Stigma. https://doi.org/10.3390/bs15121724

Serena Jinchen Xie. (2024). Cultural Prompting Improves the Empathy and Cultural Responsiveness of GPT-Generated Therapy Responses. https://pmc.ncbi.nlm.nih.gov/articles/PMC12919561/

Shah S. (2026). Substance-induced manic psychosis in which delusions were corroborated by a chatbot - case report. https://doi.org/10.21203/rs.3.rs-8919841/v1

Shayla Sharmin. (2026). Beyond Judgment: Exploring Large Language Models as Non-Judgmental Support for Maternal Mental Health.

Siddals S. (2024). "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. https://doi.org/10.1038/s44184-024-00097-4

Sohhyung Park. (2026). PsyProbe: Proactive and Interpretable Dialogue through User State Modeling for Exploratory Counseling.

Sontaga G. Forane. (2026). An Ubuntu-Guided Large Language Model Framework for Cognitive Behavioral Mental Health Dialogue. https://doi.org/10.48550/arxiv.2601.06875

Sophie Dallison. (2025). Using Generative AI to Co-Design Digital Mental Health Interventions With Adolescents in Rural South Africa: Qualitative Thematic Analysis of Participatory Workshops. https://doi.org/10.2196/73535

Stav Yosef. (2025). The impact of fine-tuning LLMs on the quality of automated therapy assessed by digital patients. https://doi.org/10.1038/s44184-025-00159-1

Steven Siddals. (2024). "It just happened to be the perfect thing": Real-life experiences of generative AI chatbots for mental health. https://doi.org/10.21203/rs.3.rs-4612612/v1

Sufyan NS. (2024). Artificial intelligence and social intelligence: preliminary comparison study between AI models and psychologists. https://doi.org/10.3389/fpsyg.2024.1353022

Suhas BN. (2026). AI Safety Training Can be Clinically Harmful. https://doi.org/10.48550/arxiv.2604.23445

Sujeong Seo. (2025). Performance Assessment of Large Language Models in Medical Consultation: Comparative Study. https://doi.org/10.2196/64318

Sunju Lee. (2025). A Study on AI Applications for Mental Health Crisis Response in Gyeonggi-do. https://doi.org/10.21032/jhis.2025.50.4.401

Suyeon Lee. (2024). Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory. https://doi.org/10.48550/arxiv.2407.03103

Till Scholich. (2025). A Comparison of Responses from Human Therapists and Large Language Model-Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. https://doi.org/10.2196/69709

Tine Kolenik. (2024). Computational Psychotherapy System for Mental Health Prediction and Behavior Change with a Conversational Agent. https://doi.org/10.2147/ndt.s417695

Tomoko Kishimoto. (2025). Single online self-compassion writing intervention reduces anxiety: With the feedback of ChatGPT. https://doi.org/10.1016/j.invent.2025.100810

Truong Le Minh Toan. (2026). Can Virtual Agents Care? Designing an Empathetic and Personalized LLM-Driven Conversational Agent.

U Selen Kilic. (2026). Integrating Artificial Intelligence Into Exposure Therapy: A One Year Follow-Up Case Report of Emetophobia With Comorbid Panic Disorder. https://doi.org/10.1002/ccr3.71715

Uscher-Pines L. (2026). Assessing Generative AI Chatbots for Alcohol Misuse Support: A Longitudinal Simulation Study. https://doi.org/10.1056/aics2500676

Vera Békés. (2026). Who Wants to Have an AI Therapist? Acceptance of Using Artificial Intelligence for Mental Health Interventions Among Clinicians, Patients and the General Community. https://doi.org/10.1002/cpp.70220

Viet Cuong Nguyen. (2026). CALM-IT: Generating Realistic Long-Form Motivational Interviewing Dialogues with Dual-Actor Conversational Dynamics Tracking.

Vowels LM. (2024). Evaluating the Efficacy of Amanda: A Voice-Based Large Language Model Chatbot for Relationship Challenges. https://doi.org/10.31234/osf.io/3x7e8

Vowels LM. (2025). The efficacy, feasibility, and technical outcomes of a GPT-4o-based chatbot Amanda for relationship support: A randomized controlled trial. https://doi.org/10.1371/journal.pmen.0000411

W Pichowicz. (2025). Performance of mental health chatbot agents in detecting and managing suicidal ideation. https://doi.org/10.1038/s41598-025-17242-4

Wang Y. (2024). Chatbot-Based Interventions for Mental Health Support. https://doi.org/10.31234/osf.io/xj7cz

Wen Zhang. (2026). GPT-Powered Chatbot-Based Positive Psychology Intervention for Well-Being Among Parents of Children With Autism Spectrum Disorder: Single-Arm Mixed Methods Study. https://doi.org/10.2196/85060

Wing Man Keung. (2025). Attitudes towards AI counseling: the existence of perceptual fear in affecting perceived chatbot support quality. https://doi.org/10.3389/fpsyg.2025.1538387

Xianrong Yao. (2025). Empathy-R1: A Chain-of-Empathy and Reinforcement Learning Framework for Long-Form Mental Health Support. https://doi.org/10.48550/arxiv.2509.14851

Xiaochen Luo. (2025). Seeking Emotional and Mental Health Support From Generative AI: Mixed-Methods Study of ChatGPT User Experiences. https://doi.org/10.2196/77951

Xiaojie Peng. (2025). Fostering adolescent engagement in generative AI art therapy: a dual SEM-ANN analysis of emotional. https://doi.org/10.3389/fpsyg.2025.1628471

Xiaoli Wu. (2025). Trust, Anxious Attachment, and Conversational AI Adoption Intentions in Digital Counseling: A Preliminary Cross-Sectional Questionnaire Study. https://doi.org/10.2196/68960

Xiaolu Dai. (2025). The paradox of agency in psychotherapy: How people with mental distress experience support from generative AI chatbots and human therapists. https://doi.org/10.1186/s12888-025-07671-w

Xiaoyi Wang. (2025). Feel the Difference? A Comparative Analysis of Emotional Arcs in Real and LLM-Generated CBT Sessions. https://doi.org/10.48550/arxiv.2508.20764

Xuan Zhang. (2025). Exploring Body Image Awareness With a Large Language Model-Based Conversational Agent: Qualitative Study With Young Adults. https://doi.org/10.2196/78829

Xueting Cui. (2025). Development and evaluation of LLM-based suicide intervention chatbot. https://doi.org/10.3389/fpsyt.2025.1634714

Xueying Bao. (2025). eCBT-I dialogue system: a comparative evaluation of large language models and adaptation strategies for insomnia treatment. https://doi.org/10.1186/s12967-025-06871-y

Yejin Kim. (2025). Aligning large language models for cognitive behavioral therapy: a proof-of-concept study. https://doi.org/10.3389/fpsyt.2025.1583739

Yeo YH. (2025). Evaluating for Evidence of Sociodemographic Bias in Conversational AI for Mental Health Support. https://doi.org/10.1089/cyber.2024.0199

Yihe Zhang. (2026). MHDash: An Online Platform for Benchmarking Mental Health-Aware AI Assistants.

Yimeng Wang. (2026). Exploring Customizable Interactive Tools for Therapeutic Homework Support in Mental Health Counseling.

You C. (2025). Alter egos alter engagement: perspective-taking can improve disclosure quantity and depth to AI chatbots in promoting mental wellbeing. https://doi.org/10.3389/fdgth.2025.1655860

Yougen Zhou. (2026). PRMB: Benchmarking Reward Models in Long-Horizon CBT-based Counseling Dialogue. https://doi.org/10.48550/arxiv.2603.11494

Youyou Cheng. (2026). The Slow Drift of Support: Boundary Failures in Multi-Turn Mental Health LLM Dialogues.

Yuval Haber. (2025). The externalization of internal experiences in psychotherapy through generative artificial intelligence: a theoretical, clinical, and ethical analysis. https://doi.org/10.3389/fdgth.2025.1512273

Yuval Haber. (2025). Validating GenAI feedback in suicide prevention training: a mixed-methods study of QPR skill assessment. https://doi.org/10.3389/fmed.2025.1709743

Zhi Liu. (2026). A fine-tuned large language model chatbot for multi-scenario radiology cancer care: randomized controlled trial on interaction optimization, emotional support, and provider burnout reduction. https://doi.org/10.1186/s12967-026-07738-6

Zhijun Guo. (2024). Evaluating the Feasibility and Acceptability of a GPT-Based Chatbot for Depression Screening: A Mixed-Methods Study. https://doi.org/10.1007/978-3-031-67278-1_20

Zhou. (2025). DiaCBT: A Long-Periodic Dialogue Corpus Guided by Cognitive Conceptualization Diagram for CBT-based Psychological Counseling. https://doi.org/10.48550/arxiv.2509.02999

Zilin Ma. (2023). Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. https://doi.org/10.48550/arxiv.2307.15810

Zixin Xiong. (2026). TrustMH-Bench: A Comprehensive Benchmark for Evaluating the Trustworthiness of Large Language Models in Mental Health.

Zoren Christian Magaan. (2026). PERPY: A Guidance Counselor Support Tool for Student Assessment and Well-Being Enhancement Using Natural Language Understanding. https://doi.org/10.36948/ijfmr.2026.v08i02.71447

Referências contextuais

Autor não informado. (2024). ACCEPTABILITY OF A CHATBOT FOR INFORMATION AND ADVICE ON DEMENTIA CAREGIVING. https://pmc.ncbi.nlm.nih.gov/articles/PMC11692751/

Autor não informado. (2025). SUN-806 Do Customized Chatbots Help with Diabetes Management? https://pmc.ncbi.nlm.nih.gov/articles/PMC12543897/

Autor não informado. (2025). The Structure of Major Life Transitions Among Older Suicide Decedents: An Application of Large Language Models. https://pmc.ncbi.nlm.nih.gov/articles/PMC12762265/

Acacia Parks. (2025). Is This Chatbot Safe and Evidence-Based? A Call for the Critical Evaluation of Generative AI Mental Health Chatbots. https://doi.org/10.2196/69534

Adhikary PK. (2024). Exploring the Efficacy of Large Language Models in Summarizing Mental Health Counseling Sessions: Benchmark Study. https://doi.org/10.2196/57306

Akhil P. Joseph. (2025). Redefining communication in mental healthcare: generative AI for neurodivergent equity and non-verbal autistic inclusion. https://doi.org/10.3389/fpsyt.2025.1611101

Alan C Y Tong. (2025). Effectiveness of Topic-Based Chatbots on Mental Health Self-Care and Mental Well-Being: Randomized Controlled Trial. https://doi.org/10.2196/70436

Albanese J. (2026). From simulated empathy to structural attunement: Realtime Editable Memory Topology and the evolution of emotionally grounded AI. https://doi.org/10.3389/frai.2026.1749517

Alexander Vanin. (2025). Psychotherapist remarks' ML classifier: insights from LLM and topic modeling application. https://doi.org/10.3389/fpsyt.2025.1608163

Alexandre Hudon. (2025). Delusional Experiences Emerging From AI Chatbot Interactions or "AI Psychosis". https://doi.org/10.2196/85799

Alifya Mukadam. (2025). Beyond Traditional Simulation: An Exploratory Study on the Effectiveness and Acceptability of ChatGPT-4o Advanced Voice Mode for Communication Skills Practice Among Medical Students. https://doi.org/10.7759/cureus.84381

Allison Diane Ihle. (2026). Training an AI Chatbot to Manage Health in Underserved Populations: Methodological Approach. https://doi.org/10.2196/84145

Amala Arul Malar Umakanth. (2025). Conversational GenAI agents in mobile health and fitness apps. https://doi.org/10.30574/wjaets.2025.15.3.1100

Amit Baumel. (2025). More than a chatbot: a practical framework to harness artificial intelligence across key components to boost digital therapeutics quality. https://doi.org/10.3389/fdgth.2025.1541676

Amos Grünebaum. (2025). Generative artificial intelligence for counseling of fetal malformations following ultrasound diagnosis. https://doi.org/10.1515/jpm-2025-0367

Angela Chen. (2026). Empirical Modeling of Therapist-Client Dynamics in Psychotherapy Using LLM-Based Assessments.

Anqi Li. (2026). CARE: An Explainable Computational Framework for Assessing Client-Perceived Therapeutic Alliance Using Large Language Models.

Anqi Li. (2026). RECAP: Resistance Capture in Text-based Mental Health Counseling with Large Language Models.

Anthony Mina. (2025). AI chatbots as 'pocket doctors': intimate health support for young women in Lebanon. https://doi.org/10.1186/s12889-025-25386-1

Antonio Carnevale. (2025). Empatia artificiale e dialettiche della vulnerabilità. Ripensare la psicoterapia nell’era digitale. https://doi.org/10.15162/1827-5133/2237

Arya Rao. (2025). A Future of Self-Directed Patient Internet Research: Large Language Model-Based Tools Versus Standard Search Engines. https://doi.org/10.1007/s10439-025-03701-6

Aya Elsayed Abdelwahed. (2025). Public attitudes and practices toward using AI chatbots for healthcare assistance: a multinational cross-sectional study. https://doi.org/10.1186/s12913-025-13832-0

Baihan Lin. (2025). COMPASS: Computational mapping of patient-therapist alliance strategies with language modeling. https://doi.org/10.1038/s41398-025-03379-3

Benjamin kate. (2024). The Role of AI in Mental Health: Breaking Down Barriers for Men. https://doi.org/10.20944/preprints202411.0007.v1

Benjamin Krichevsky. (2026). Human vs. artificial intelligence: Physicians outperform ChatGPT in real-world pharmacotherapy counselling. https://doi.org/10.1002/bcp.70321

Bijoyaa Mohapatra. (2026). Assistive Intelligence: A Framework for AI-Powered Technologies Across the Dementia Continuum. https://doi.org/10.3390/jal6010008

Blease C. (2023). ChatGPT and mental healthcare: balancing benefits with risks of harms. https://doi.org/10.1136/bmjment-2023-300884

Bosco Garcia. (2025). The problem of atypicality in LLM-powered psychiatry. https://doi.org/10.1136/jme-2025-110972

Brian Yim. (2026). Exploring the application of large language models in coding the experiencing scale (EXP). https://doi.org/10.1080/28324765.2026.2664163

Chihying Liao. (2024). AI-Enhanced Healthcare: Integrating ChatGPT-4 in ePROs for Improved Oncology Care and Decision-Making: A Pilot Evaluation. https://doi.org/10.3390/curroncol32010007

Chit Thet Lal Oo. (2026). The benefits and future potential of generative artificial intelligence (GAI) on mental health: a Delphi study. https://doi.org/10.1080/17482631.2026.2621802

Chung Man Ho. (2025). Development and Validation of a Large Language Model-Powered Chatbot for Neurosurgery: Mixed Methods Study on Enhancing Perioperative Patient Education. https://doi.org/10.2196/74299

Craig W McFarland. (2026). Advancing neurotech justice in youth digital mental health: insights from an interdisciplinary and cross-generational workshop. https://doi.org/10.1038/s44277-025-00052-x

Crawford A. (2026). Urgent considerations for suicide prevention in the safe and ethical use of artificial intelligence. https://doi.org/10.1503/cmaj.251693

D. Lozoya. (2024). Generating Mental Health Transcripts with SAPE (Spanish Adaptive Prompt Engineering) https://doi.org/10.18653/v1/2024.naacl-long.285

Daeun Lee. (2026). Before and After ChatGPT: Revisiting AI-Based Dialogue Systems for Emotional Support.

Daniel Cabrera Lozoya. (2025). Leveraging Large Language Models for Simulated Psychotherapy Client Interactions: Development and Usability Study of Client101. https://doi.org/10.2196/68056

Daniel Szoke. (2025). Artificial Intelligence in Mental Health Services Under Illinois Public Act 104-0054: Legal Boundaries and a Framework for Establishing Safe, Effective AI Tools. https://doi.org/10.2196/84854

Deepak Kumar. (2026). Artificial Intelligence and Mental Health: Assessing Its Impact on Psychological Well-Being Among Indian Youth. https://doi.org/10.33545/26648903.2026.v8.i4a.173

Denecke K. (2024). The ethical aspects of integrating sentiment and emotion analysis in chatbots for depression intervention. https://doi.org/10.3389/fpsyt.2024.1462083

Dominik Wolff. (2025). Personalized Support in Hereditary Breast and Ovarian Cancer After Genetic Counseling by the Chatbot-Based GENIE Mobile App: Proof-of-Concept Wizard of Oz Study. https://doi.org/10.2196/69115

DR. P. RAMESH BABU. (2026). GENERATIVE AI AGENT EMPOWERING YOUTH MENTAL HEALTH. https://doi.org/10.5281/zenodo.19159921

Drin Ferizaj. (2025). Identifying Yalom's group therapeutic factors in anonymous mental health discussions on Reddit: a mixed-methods analysis using large language models, topic modeling and human supervision. https://doi.org/10.3389/fpsyt.2025.1503427

Dujrudee Chinwong. (2025). Integrating ChatGPT for smoking cessation counseling practice in pharmacy education: A single group quasi-experimental study. https://doi.org/10.18332/tid/211706

Elena Markova. (2026). Embodied simulation, body language, and symbolization: understanding somatic symptoms in psychoanalysis. https://doi.org/10.3389/fpsyg.2026.1642418

Elizabeth Ajalo. (2025). Widespread use of ChatGPT and other Artificial Intelligence tools among medical students in Uganda: A cross-sectional study. https://doi.org/10.1371/journal.pone.0313776

Emily Saltz. (2026). AI and Suicide Prevention: A Cross-Sector Primer. https://doi.org/10.48550/arxiv.2605.04321

Esteban Zavaleta-Monestel. (2026). When Compliance Is Not Safety: The Regulatory Blind Spot in AI Companion Chatbots. https://doi.org/10.7759/cureus.105902

Evan Matthews. (2026). Considerations about the proliferation of large language model chatbots and youth mental health. https://doi.org/10.1017/ipm.2026.10195

Fabio Frisone. (2026). Toward clinical integration of generative AI in mental health: personalization, multimodality and inter-entity experience. https://doi.org/10.3389/fpubh.2026.1603238

Ferrario A. (2024). The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals With Depression: A Critical Analysis. https://doi.org/10.2196/56569

Figueroa RA. (2026). Towards scalable psychological first aid training: an autoethnographic exploration of the potential of large language models in simulation-based learning. https://doi.org/10.1080/20008066.2026.2616976

Flavia Morfini. (2025). The Autism Open Clinical Model (A.-O.C.M.) as a Phenomenological Framework for Prompt Design in Parent Training for Autism: Integrating Embodied Cognition and Artificial Intelligence. https://doi.org/10.3390/brainsci15111213

Francis C Ohu. (2025). Public Health Risk Management, Policy, and Ethical Imperatives in the Use of AI Tools for Mental Health Therapy. https://doi.org/10.3390/healthcare13212721

Fujita J. (2024). Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: A Randomized Control Study Termination Report. https://doi.org/10.1101/2024.11.25.24317880

Giorgia Pozzi. (2025). Keeping an AI on the mental health of vulnerable populations: reflections on the potential for participatory injustice. https://doi.org/10.1007/s43681-024-00523-5

Haber Y. (2024). The Artificial Third: A Broad View of the Effects of Introducing Generative Artificial Intelligence on Psychotherapy. https://doi.org/10.2196/54781

Hamilton Morrin. (2026). It Is the Journey, Not the Destination: Moving From End Points to Trajectories When Assessing Chatbot Mental Health Safety. https://doi.org/10.2196/91454

Han GI. (2026). Advantages and challenges for utilization of generative artificial intelligence in clinical nursing practice: an integrative review. https://doi.org/10.1186/s12912-026-04518-x

Hannah van Kolfschooten. (2025). AI chatbots for promoting healthy habits: Legal, ethical, and societal considerations. https://doi.org/10.1177/20552076251390004

Hassan Alhuzali. (2025). Pre- Trained Language Models for Mental Health: An Empirical Study on Arabic Q&A Classification. https://doi.org/10.3390/healthcare13090985

Hayato Ebihara. (2025). Development of a Clinical Clerkship Mentor Using Generative AI and Evaluation of Its Effectiveness in a Medical Student Trial Compared to Student Mentors: 2-Part Comparative Study. https://doi.org/10.2196/76702

Held P. (2025). Generative artificial intelligence in posttraumatic stress disorder treatment: Exploring five different use cases. https://doi.org/10.1002/jts.23188

Hessah Al Suwaidan. (2026). Comparing Perceptions of ChatGPT Use in Health Attitude Contexts Among Users and Nonusers: Cross-Sectional Study. https://doi.org/10.2196/79276

Hilal Toklu Baloğlu. (2025). Effect of ChatGPT use on eating disorders and body image. https://doi.org/10.5498/wjp.v15.i8.107122

İçen S. (2025). Artificial intelligence guidance in ethically challenging clinical scenarios in child and adolescent psychiatry: a qualitative study in the context of Turkiye. https://doi.org/10.1186/s12910-025-01323-0

J. Torous. (2023). Focusing on Digital Research Priorities for Advancing the Access and Quality of Mental Health. https://doi.org/10.2196/47898

Janett V Chavez Sosa. (2026). Anxiety and Depression Associated With the Dependent Use of Generative AI in Medical Students: Cross-Sectional Study. https://doi.org/10.2196/82667

Jean-Christophe Bélisle-Pipon. (2026). Fatal deception: how generative AI fosters therapeutic misconception in vulnerable users. https://doi.org/10.3389/fdgth.2026.1756620

Jesudason D. (2025). Artificial intelligence (AI) in psychotherapy: A challenging frontier. https://doi.org/10.1177/10398562251346075

Jianchen Luo. (2026). Optimization of University Counseling Consent Forms With Large Language Models: Multidimensional Comparative Evaluation. https://doi.org/10.2196/86502

Jiang Ji. (2026). Comparison of online radiologists and large language model chatbots in responding to common radiology-related questions in Chinese: a cross-sectional comparative analysis. https://doi.org/10.21037/qims-2025-1716

Jiaqing Wang. (2026). Elder-Sim: A Psychometrically Validated Platform for Personality-Stable Elderly Digital Twins. https://doi.org/10.64898/2026.03.25.26349036

Jonathan A Tangsrivimol. (2025). Benefits, limits, and risks of ChatGPT in medicine. https://doi.org/10.3389/frai.2025.1518049

José Ganicho. (2025). Use of ChatGPT in HIV Infection Counselling and Literacy. https://doi.org/10.20344/amp.22805

Junichi Fujita. (2025). Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report. https://doi.org/10.2196/70960

Jurblum M. (2025). Potential promises and perils of artificial intelligence in psychotherapy -The AI Psychotherapist (APT). https://doi.org/10.1177/10398562241286312

Kalam KT. (2024). ChatGPT and mental health: Friends or foes? https://doi.org/10.1002/hsr2.1912

Karen Yirmiya. (2025). Mentalizing Without a Mind: Psychotherapeutic Potential of Generative AI. https://doi.org/10.2196/79156

Karin Hammerfald. (2025). Leveraging large language models to identify microcounseling skills in psychotherapy transcripts. https://doi.org/10.1080/10503307.2025.2539405

Kayleigh-Ann Clegg. (2025). Shoggoths, Sycophancy, Psychosis, Oh My: Rethinking Large Language Model Use and Safety. https://doi.org/10.2196/87367

Ke Zhang. (2025). Effects of attractions and social attributes on peoples' usage intention and media dependence towards chatbot: The mediating role of parasocial interaction and emotional support. https://doi.org/10.1186/s40359-025-03284-w

Keshavan M. (2026). Do generative AI chatbots increase psychosis risk? https://doi.org/10.1002/wps.70017

Khan N. (2026). Bridging Science and Subjectivity: Evolving Evidence, Emerging Technologies and the Call for Personalised Psychotherapy: Créer une passerelle entre la science et la subjectivité : Évolution des données probantes, arrivée de nouvelles technologies et appel à une psychothérapie personnalisée. https://doi.org/10.1177/07067437261425840

Kuhlmeier FO. (2025). Designing Chatbots to Treat Depression in Youth: Qualitative Study. https://doi.org/10.2196/66632

Kunal A. Sapkale. (2026). HearU: A Smart AI-Based Mental Health Chatbot Using NLP for Emotional Support and Therapy. https://doi.org/10.48175/ijarsct-32305

Lawrence HR. (2024). The Opportunities and Risks of Large Language Models in Mental Health. https://doi.org/10.2196/59479

Lei Xian. (2025). My digital mentor: a mixed-methods study of user-GAI interactions. https://doi.org/10.3389/fpsyg.2025.1636480

Li S. (2025). Determinants of rural middle school students' adoption of AI chatbots for mental health. https://doi.org/10.3389/fpubh.2025.1619535

Linfeng Wang. (2025). Application of Narrative and AI-Assisted Follow-Up After Voluntary Medical Male Circumcision: Multicenter, Double-Blind, Prospective, Randomized Controlled Trial. https://doi.org/10.2196/68573

Lingling Xu. (2025). Personalized rTMS treatment recommendation with retrieval-augmented LLM reasoning. https://doi.org/10.1186/s40708-025-00275-w

Liu Lingyun. (2025). CONSTRUCTION AND PRACTICAL EXPLORATION OF AN AI EMOTIONAL COMPANION MODEL FOR COLLEGE STUDENTS. https://doi.org/10.5281/zenodo.18023370

Lizhen Lu. (2025). Healthcare professionals and the public sentiment analysis of ChatGPT in clinical practice. https://doi.org/10.1038/s41598-024-84512-y

Malhotra C. (2025). Supporting Dementia Caregiving With a Mobile Care Ecosystem: Development and Mixed Methods Study. https://doi.org/10.2196/78759

Maria Cecilia Lopes. (2025). The future of the sleep field using large language models in mental health care. https://doi.org/10.5664/jcsm.11766

Mariusz Panczyk. (2026). Effect of Emotional Prompt on the Quality of ChatGPT Based Discharge Instructions After Laparoscopic Cholecystectomy Using ERAS Derived Benchmarks. https://doi.org/10.2147/jmdh.s566966

Max Ostermann. (2025). If a therapy bot walks like a duck and talks like a duck then it is a medically regulated duck. https://doi.org/10.1038/s41746-025-02175-z

Mehmet Akkurt. (2025). Learning Through Simulation: Counselor Trainees' Interactions with ChatGPT as a Client. https://doi.org/10.3390/bs15121660

Michael R MacIntyre. (2026). When the therapist hallucinates: AI, psychedelics, and the risks of unsupervised digital mental health care. https://doi.org/10.1177/20552076261429341

Mingjun Zhang. (2025). Performance of Large Language Models in Chinese Language Medical Counseling on. https://doi.org/10.2147/idr.s553523

Mostafa Abdou. (2025). Leveraging Large Language Models to Estimate Clinically Relevant Psychological Constructs in Psychotherapy Transcripts. https://doi.org/10.5334/cpsy.141

Muhammet Hüseyin Erkan. (2026). ChatGPT and Gemini in warfarin counseling. https://doi.org/10.3325/cmj.2025.66.399

Muthukumar K. (2025). Empathy AI in healthcare. https://doi.org/10.3389/fpsyg.2025.1680552

Nan Bai. (2025). Detecting Sociodemographic Biases in the Content and Quality of Large Language Model-Generated Nursing Care: Cross-Sectional Simulation Study. https://doi.org/10.2196/78132

Neary M. (2025). Think FAST: a novel framework to evaluate fidelity, accuracy, safety, and tone in conversational AI health coach dialogues. https://doi.org/10.3389/fdgth.2025.1460236

Nick Kabrel. (2025). When can AI psychotherapy be considered comparable to human psychotherapy? Exploring the criteria. https://doi.org/10.3389/fpsyt.2025.1674104

Nithesh K. (2026). Generative AI for Mental Health: Bridging Mental Health Gaps with An AI-Driven Emotional Support System. https://doi.org/10.47392/irjaeh.2026.0271

Olla P. (2025). Beyond the Bot: A Dual-Phase Framework for Evaluating AI Chatbot Simulations in Nursing Education. https://doi.org/10.3390/nursrep15080280

Oscar Robayo-Pinzon. (2025). Generative artificial intelligence (GenAI) use and dependence: an approach from behavioral economics. https://doi.org/10.3389/fpubh.2025.1634121

Pablo Roca. (2026). Artificial intelligence in the psychologist's toolkit: Psypilot as a case study. https://doi.org/10.3389/fpsyg.2026.1775464

Patrick Baxter. (2025). Public Versus Academic Discourse on ChatGPT in Health Care: Mixed Methods Study. https://doi.org/10.2196/64509

Paz Mor Naim. (2025). Preprocessing Large-Scale Conversational Datasets: A Framework and Its Application to Behavioral Health Transcripts. https://doi.org/10.2196/78082

Pearla Papiernik. (2025). Acceptability of a Conversational Agent-Led Digital Program for Anxiety: Mixed Methods Study of User Perspectives. https://doi.org/10.2196/76377

Priyanka Singh. (2025). Adaptive LLM Agents: Toward Personalized Empathetic Care.

Qing Han. (2025). Unleashing the potential of chatbots in mental health: bibliometric analysis. https://doi.org/10.3389/fpsyt.2025.1494355

Raghu Chukkala. (2025). Beyond Conversational Interfaces: The Emergence of Cognitive Conversational AI (CCAI) and Its Role in Redefining HumanMachine Symbiosis. https://doi.org/10.35629/5252-0705351359

Ragozzino O. (2025). Inclusivity and Innovation: A Pilot Study on Free AI Models for Clinical Supervision in Gestalt Therapy (GT) https://doi.org/10.20944/preprints202510.0727.v1

Ruiyi Wang. (2024). PATIENT-\psi: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. https://doi.org/10.48550/arxiv.2405.19660

Salman Bukhari. (2025). Wearable IoMT, Edge AI and LLMs for Monitoring Health and Wellness: A Harmonization Framework for Responsible AI Interventions. https://doi.org/10.33774/coe-2025-1cnc9

Santosh Patapati. (2025). A Framework for ECA-Based Psychotherapy. https://doi.org/10.31224/4977

Sarah Ying Tse Tan. (2025). Accuracy of Large Language Model Responses Versus Internet Searches for Common Questions About Glucagon-Like Peptide-1 Receptor Agonist Therapy: Exploratory Simulation Study. https://doi.org/10.2196/78289

Sarkar S. (2023). A review of the explainability and safety of conversational agents for mental health to identify avenues for improvement. https://doi.org/10.3389/frai.2023.1229805

Savio A. (2026). A review of artificial intelligence enhanced cognitive behavioural therapy using the BECK AI BOT for mental health interventions. https://doi.org/10.1007/s44192-026-00391-x

Sebastian Dohnány. (2026). Technological. https://doi.org/10.1038/s44220-026-00595-8

Sergei Koltcov. (2024). Using large language models for extracting and pre-annotating texts on mental health from noisy data in a low-resource language. https://doi.org/10.7717/peerj-cs.2395

Shane Cross. (2025). Insights from fifteen years of real-world development, testing and implementation of youth digital mental health interventions. https://doi.org/10.1016/j.invent.2025.100849

Sharon Grundmann. (2025). Lilobot: A Cognitive Conversational Agent to Train Counsellors at Children's Helplines : Design and Initial Evaluation. https://doi.org/10.1007/s10916-024-02121-8

Shengyu He. (2026). Power Distance and Psychological Safety in LLM Counseling: Effects on Self-Efficacy with Implications for Mental Health-Relevant Behavior Change. https://doi.org/10.3390/bs16020241

Shubham Ravindra Sali. (2025). AI-Driven Mental Health Apps, Chatbots, and Teletherapy Platforms Transforming Therapy Accessibility. https://doi.org/10.63096/medtigo3084233

Simelane PM. (2026). Leveraging chatbots for enhanced decision-making: a comprehensive literature review. https://doi.org/10.3389/frai.2026.1748544

Simone Schmidt. (2025). Psychology student and mental health practitioner experiences of and perspectives on Client101, a virtual client chatbot training tool. https://doi.org/10.1186/s12909-025-07668-9

Siyeon Ko. (2025). Users' Needs for Mental Health Apps: Quality Evaluation Using the User Version of the Mobile Application Rating Scale. https://doi.org/10.2196/64622

Sorio Boit. (2025). A Prompt Engineering Framework for Large Language Model-Based Mental Health Chatbots: Conceptual Framework. https://doi.org/10.2196/75078

Stade EC. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. https://doi.org/10.1038/s44184-024-00056-z

Stefano Valente. (2026). Large Language Models as Digital Philotherapy A Low-Intensity, Non-Substitutive Framework for Mental Health Support in Contexts of Limited Access. https://doi.org/10.5281/zenodo.18385137

Steffi G Riedel-Heller. (2026). [The Silent Risk: Loneliness and Health]. https://doi.org/10.1055/a-2776-3877

Steinbrenner T. (2025). Exploring negative experiences in psychotherapy: Patient dissatisfaction as a use case for a new NLP approach leveraging big data from online forums. https://doi.org/10.21203/rs.3.rs-6766808/v1

Sultan Alam. (2024). Integrating ChatGPT: Enhancing postpartum mental healthcare with artificial intelligence (AI) support. https://doi.org/10.1177/20552076241295565

Sun X. (2026). AI companions and adolescent social relationships: Benefits, risks, and bidirectional influences. https://doi.org/10.1093/cdpers/aadaf009

Suzanne Spencer. (2025). Enabling access or automating empathy? Using chatbots to support GBV survivors in conflicts and humanitarian emergencies. https://doi.org/10.1017/s1816383125100763

Tafani. (2025). A chi servono le macchine per scrivere frasi probabili. https://doi.org/10.5281/zenodo.17814686

Tao Li. (2025). THE RESHAPING OF UNIVERSITY STUDENT AFFAIRS MANAGEMENT MECHANISM BASED ON ARTIFICIAL INTELLIGENCE TECHNOLOGY: INTEGRATION OF EDUCATIONAL THEORY AND CONSTRUCTION OF PSYCHOLOGICAL SUPPORT PATHWAYS. https://doi.org/10.35631/ijepc.1060032

Thomas Kallstenius. (2025). Comparing traditional natural language processing and large language models for mental health status classification: a multi-model evaluation. https://doi.org/10.1038/s41598-025-08031-0

Tobias Steinbrenner. (2025). Exploring negative experiences in psychotherapy using an NLP approach on online forum data. https://doi.org/10.1038/s44184-025-00172-4

Tony Rousmaniere. (2026). Large language models as mental health providers. https://doi.org/10.1016/s2215-0366(25)00269-x

Vaishali Katti. (2026). Personalized Depression Management System Using LLMs and Reinforcement Learning: A Survey. https://doi.org/10.1007/978-981-95-5079-1_2

van der Vet I. (2025). Copilot in service: Exploring the potential of the large language model-based chatbots for fostering evaluation culture in preventing and countering violent extremism. https://doi.org/10.12688/openreseurope.19612.2

Vinh T. (2026). Psychiatry's Blind Spot: Independent Use of General-Purpose Large Language Models by Individuals With Psychopathology. https://doi.org/10.1016/j.mcpdig.2026.100353

Vivian Hui. (2026). Perspectives and preferences of domestic violence survivors regarding digital platform and AI chatbot for help-seeking: A qualitative study. https://doi.org/10.1371/journal.pone.0342453

Wang E. (2025). Adoption and perception of LLM-based chatbots in health care: an exploratory cross-sectional survey of individuals with rheumatic diseases. https://doi.org/10.1093/rap/rkaf083

Wang X. (2023). ChatGPT: promise and challenges for deployment in low- and middle-income countries. https://doi.org/10.1016/j.lanwpc.2023.100905

Wang Y. (2026). Dual Chain-Mediation of GenAI Chatbots on Loneliness: Perceived Misinformation Exposure and User Trust via Negative Emotions. https://doi.org/10.3389/ijph.2026.1609017

Wanhong Zheng. (2026). Evaluation of artificial intelligence-generated vignettes depicting patient chatbot use in psychiatric contexts. https://doi.org/10.1038/s41746-026-02605-6

Wordh Ul Hasan. (2024). Empowering Alzheimer's caregivers with conversational AI: a novel approach for enhanced communication and personalized support. https://doi.org/10.1038/s44385-024-00004-8

Wu G. (2025). Chatbots and Diabetes: Is There Gender Bias? https://doi.org/10.1177/23743735251380954

Yahui Wang. (2025). Effect of a Cognitive Behavioral Therapy-Based AI Chatbot on Depression and Loneliness in Chinese University Students: Randomized Controlled Trial With Financial Stress Moderation. https://doi.org/10.2196/63806

Yang Ni. (2025). "Even GPT Can Reject Me": Conceptualizing Abrupt Refusal Secondary Harm (ARSH) and Reimagining Psychological AI Safety with Compassionate Completion Standard (CCS)

Yanjie Chen. (2025). Are you willing to forgive generative AI doctors? Trust repair after failures in online health consultation services. https://doi.org/10.3389/fpsyg.2025.1668633

Yinru Long. (2024). AffirmativeAI: Towards LGBTQ+ Friendly Audit Frameworks for Large Language Models. https://doi.org/10.48550/arxiv.2405.04652

Yoshida I. (2026). Feasibility and User Experience of an AI-Supported mHealth Intervention for Remote Life Goal Setting Based on Flow Theory: Exploratory Within-Participant Study. https://doi.org/10.2196/78717

Yunlong Liu. (2025). AI-Enabled Personalized Smoking Cessation Intervention With the Aipaca Chatbot: Mixed Methods Feasibility Study. https://doi.org/10.2196/73319

Zhao Ni. (2025). Evaluating the Usability of an HIV Prevention Artificial Intelligence Chatbot in Malaysia: National Observational Study. https://doi.org/10.2196/70034

Zhe Gao. (2025). Bibliometric analysis of trends, innovations, and the future of CBT-based mobile interventions for depression. https://doi.org/10.3389/fmed.2025.1710291

Zhongyu Shi. (2025). Toward emotional mediation: generative AI in art therapy for psychosocial health support. https://doi.org/10.3389/fpubh.2025.1690119

Zimmerman JW. (2025). Matters arising: a response to loneliness and suicide mitigation for students using GPT3-enabled chatbots. https://doi.org/10.1038/s44184-024-00083-w

Zohar Elyospeh. (2025). The role of generative artificial intelligence in evaluating adherence to responsible press media reports on suicide: A multisite, three-language study. https://doi.org/10.1192/j.eurpsy.2025.10037

Zou. (2025). The Application of Large Language Models on Major Depressive Disorder Support Based on African Natural Products.