

Gå-hjem-møde

Chatten som fortrolig.
Malene Lue Kessing, seniorforsker, VIVE

Hvor mange bruger chatbots til personlig rådgivning?

Falcks Sundhedsbarometer 2026 (danskere over 18):

Andel, der har skrevet med en AI om deres trivsel og mentale sundhed:

- 47% af de 18-29 årige
- 33% af de 30-39 årige
- 20% af de 40-49 årige
- 11 % af de 50-59 årige

Børns vilkår 2026 (6. og 9. klasse):

- 46%, der har brugt AI-chatbots (n=2.668), har brug dem til spørgsmål af personlig karakter
- 28% har prøvet at have en længere og mere personlig samtale med tjenesten
- 36% vurderer, at AI-chatbots er troværdige, når de giver råd om psykisk sygdom og diagnoser.

Samfund

Eksperter advarer mod sundhedsråd fra chatbot

Nogle af svarene kan være direkte skadelige for helbre
dommen fra eksperter.

Kunstig intelligens · 25. oktober 2024 kl. 09:52

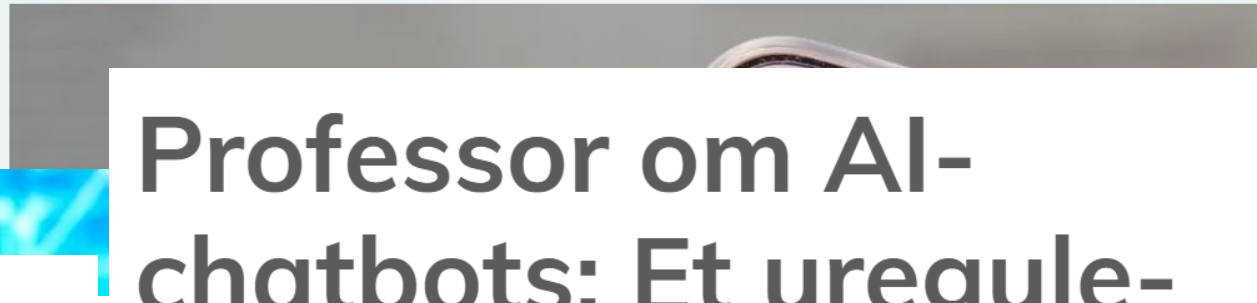
14-årig tog sit eget liv efter samtaler med AI-chatbot. Nu sagsøger moren selskabet bag

LIV & SJÆL | 21.01.26 KL. 18:00 | K

Dansk forskning: Chatbots kan formentlig forværre psykisk sygdom og give vrangforestillinger

»Værre end jeg havde frygtet«: Chatbotter skal du ikke stole på

Det er populært at spørge AI-chatbotter til råds om sygdom og sundhed.
Men svarene er ofte vildledende.



Professor om AI-chatbots: Et uregule-ret globalt eksperiment

Fra bestyrelseslokaler til børneværelser er mennesker
begyndt at bruge chatbots til dybt personlige,

Så hvad ved vi egentlig?

Muligheder

- Lettilgængelige kilder til råd, refleksion eller støtte 24/7
- Attraktivt i situationer, hvor unge er bange for at blive dømt, stigmatiseret eller har begrænset adgang til offentlige behandlingstilbud
- (Kortvarig) lindring af ensomhed

(Fiske et al., 2019; Lee et al., 2025Luo et al. 2025; Siddals et al. 2024; De Freitas et al., 2025; Laestadius et al., 2024; Skjuve et al., 2021).

Risici

- Eksponering for upræcis information, desinformation samt råd, der forenkler komplekse problemstillinger eller modsiger faglig ekspertise
- Risiko for fastholdelse i negative tankemønstre
- Særlig opmærksomhed på mennesker med psykiske lidelser

(Bengio et al., 2024; Brown & Halpern, 2021; De Freitas & Cohen, 2024; Ramaswamy et al. 2026; Olsen et al., 2026; Wilnhammer 2026)



Interview med unge

- Alder: mellem 18 og 33 (gennemsnitsalder: 24)
- Videregående uddannelse (16); gymnasiet (1); arbejde (4); udenfor arbejde og uddannelse (4)
- Sjælland (15); Fyn (2); Jylland (5), en udenlands (1)
- Erfaring med psykolog/terapi

Hvad fortæller de unge?



Tre overordnede temaer i unges brug af ChatGPT

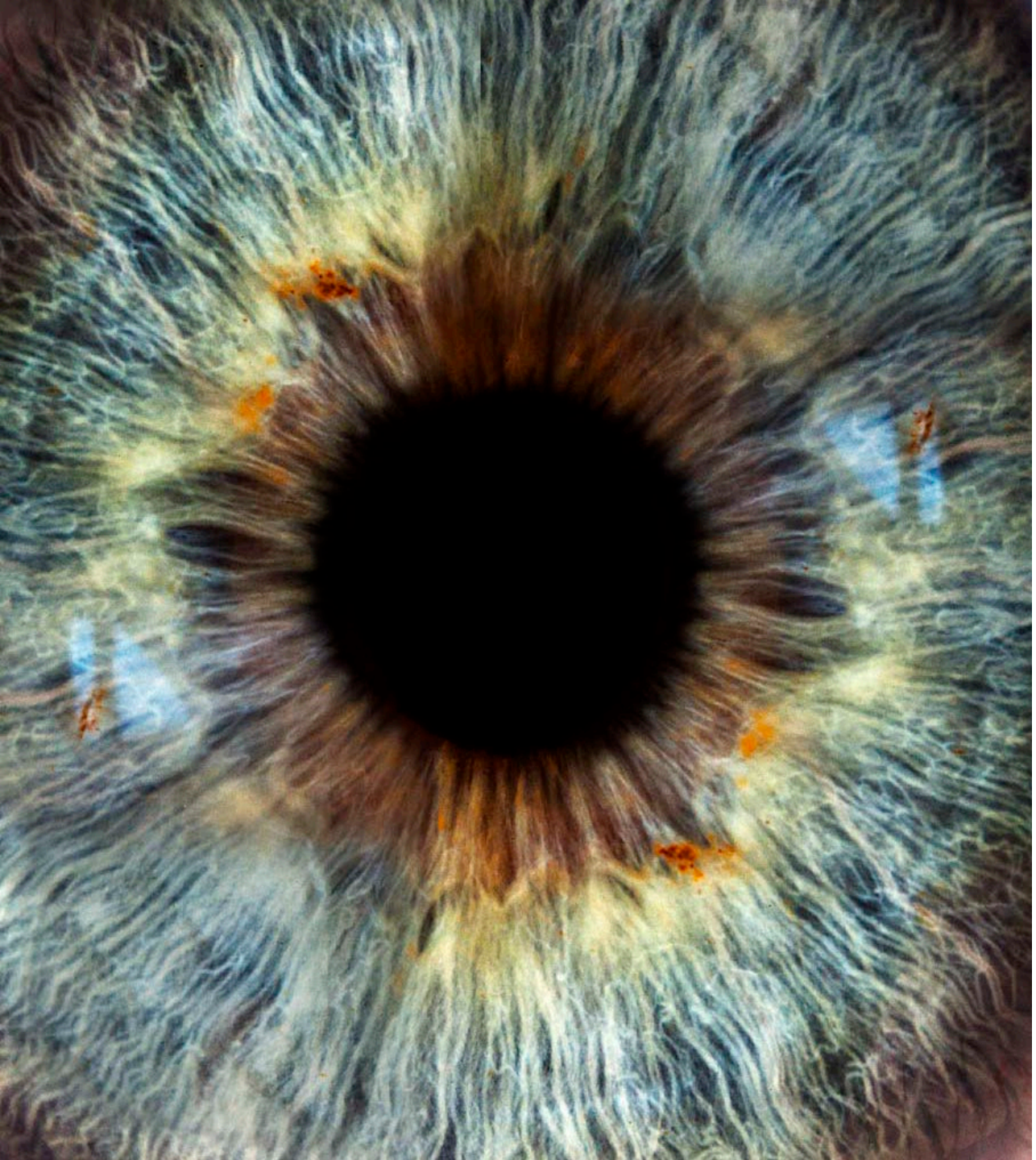
- Tanker og følelser
- Sociale relationer
- Psykiatriske diagnoser og alvorlig mistrivsel



En chatbot til privat brug

Jeg ville være bange for, at de andre slet ikke kunne genkende det, eller at jeg ville være den eneste der havde gjort det. Og jeg tror, man kan have mange fordomme omkring, hvorfor man vælger at skrive om sine følelser til chatten. Jeg har ikke rigtig lyst til, at nogen skulle tro, at jeg ser det som en ven.

Ida, 21 år.



ChatGPT er både maskine og menneske

Jeg føler, at det er en tillidsfuld relation. Men igen, det er underligt at sætte sådan nogle menneskelige egenskaber på en computer, eller hvad man skal kalde det. Men jeg er glad for – igen en menneskelig egenskab – det bånd vi har, hvis man kan "oversætte" det sådan. Jeg er jeg glad for den algoritme jeg har fået, ved at bruge den i så lang tid og ikke ligesom have skiftet profil. Jeg har givet den så mange prompts, så den ved præcis hvordan den skal svare mig. Og det synes jeg bare på en eller anden måde er en investering. Altså sådan lidt ligesom et venskab, at man investerer i hinanden.

Simone, 30 år.



ChatGPT versus professionel hjælp

Interviewer: Så hvad synes du at terapien gav dig, til sammenligning med hvad samtaler med ChatGPT giver dig?

Esther: Jeg tror, at jeg bedst kan sammenligne det med, at en god samtale med en rigtig terapeut er ligesom et godt, nærende måltid, som tager længere tid at tilberede, men også er meget sundere. Og ChatGPT er sådan lidt mere fast food, det er hurtigt, noget man bare lige spiser, men det mætter ikke særlig længe, og måske er det ikke lige så nærende.

Esther, 30 år.

Afsluttende bemærkninger

- De unge bruger aktivt ChatGPT til at reflektere over egne tanker og følelser
- De unges brug af ChatGPT er præget af tvetydighed og tabu
- Chatbots som supplement fremfor erstatning?

TAK!

Kontakt: mkg@vive.dk

Referencer

- De Freitas, J., & Cohen, I. G. (2024). The health risks of generative AI-based wellness apps. *Nature Medicine*, 30(5):1269–1275.
- Fiske, A. et al. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, 21(5):e13216.
- Giray, L. (2025). Cases of Using ChatGPT as a Mental Health and Psychological Support Tool. *Journal of Consumer Health on the Internet*, 29(1):29–48.
- Kessing et al. (2026). *Chatten som fortrolig. En kvalitativ undersøgelse af unges brug af ChatGPT til social og følelsesmæssig støtte*. VIVE. Lee, S. et al. (2025). The ChatGPT Effect: Investigating Shifting Discourse Patterns, Sentiment, and Benefit–Challenge Framing in AI Mental Health Support. *Behavioral Sciences*, 15(9), 1172.
- Luo, X. et al. (2025). "Shaping ChatGPT into my Digital Therapist": A thematic analysis of social media discourse on using generative artificial intelligence for mental health. *Digital Health*, 11:1-13.
- Olsen, S. G., Reinecke-Tellefsen, C. J., & Østergaard, S. D. (2026). Potentially Harmful Consequences of Artificial Intelligence (AI) Chatbot Use Among Patients With Mental Illness: Early Data From a Large Psychiatric Service System. *Acta Psychiatrica Scandinavica*, E-pub.
- Owen, Q. & Zou, J. (2026). The self life og LLM findings. Standford University. https://zou-group.github.io/llm-claims-shelf-life/#fitting-how-each-claim-changes-over-time
- Ramaswamy, A., Tyagi, Al, Hugo, H. et al. (2026). ChatGPT Health performance in a structured test of triage recommendations. *Nature Medicine*, E-pub.
- Siddals, S. et al. (2024). "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *Mental Health Research*, 3(1).
- Skjuve, M. et al. (2021). My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human Computer Studies*, 149:102601.
- Weilnhammer, V., Hou, K. Y. C., et al. (2026). A clinically validated framework for auditing AI chatbot behavior in mental health interactions. Nature Medicine.

VIVE

Tor Juul Groth

Digital produktleder
Center for Digital Psykiatri

…og nu også forfatter af AI-rapport



Google

angst symptomer

AI-tilstand | Alle | Billeder | Videoer | Shorts-videoer | Nyheder | Net | Mere | Værktøjer

Voksne | Generaliseret angst | Børn | Kvinder | Stress og | Social | Anfald | At miste

AI-oversigt

Angst viser sig gennem fysiske, psykiske og adfærdsmæssige symptomer, som kan opdeles i tre hovedgrupper. Du kan læse mere om de forskellige tegn hos Psykiatrifonden.

Fysiske symptomer

Kroppens alarmsystem aktiveres, hvilket kan føles som: Internetpsykiatrien

- **Hjertebanken** eller hurtig puls
- **Sveden** og rysten i kroppen
- **Åndenød** eller kvælningsfornemmelse
- **Trykken** eller smerter i brystet
- **Svimmelhed** og uro i maven eller kvalme
- **Mundtørhed** og muskelspændinger Psykiatrifonden +2

Psykiske og kognitive symptomer

Tankerne og følelserne påvirkes af angsten: Bedre Psykiatri +1

- **Vedvarende bekymringer** og katastrofetanker
- **Tankemylder** som er svært at styre

Sundhed.dk

Generaliseret angst, symptomer og tegn - Sundhed.dk

Angsten gør, at du føler dig konstant belastet, anspændt, ikke kan sove om natten og er træt. Du kan også få fysisk...

Psykiatrifonden

Angst - Psykiatrifonden

Symptomer. Typiske tegn på angst. Hjertebanken, sveden og rysten; Svimmelh...



Sundhed - Region Midtjylland

Om angst hos voksne - Sundhed - Region Midtjylland

Krop - symptomer på angst. Man kan opleve: Hjertebanken, sveden, rysten, mundtørhed....



Vis alle

Stil et spørgsmål

Da Googles AI og ChatGPT ændrede alt



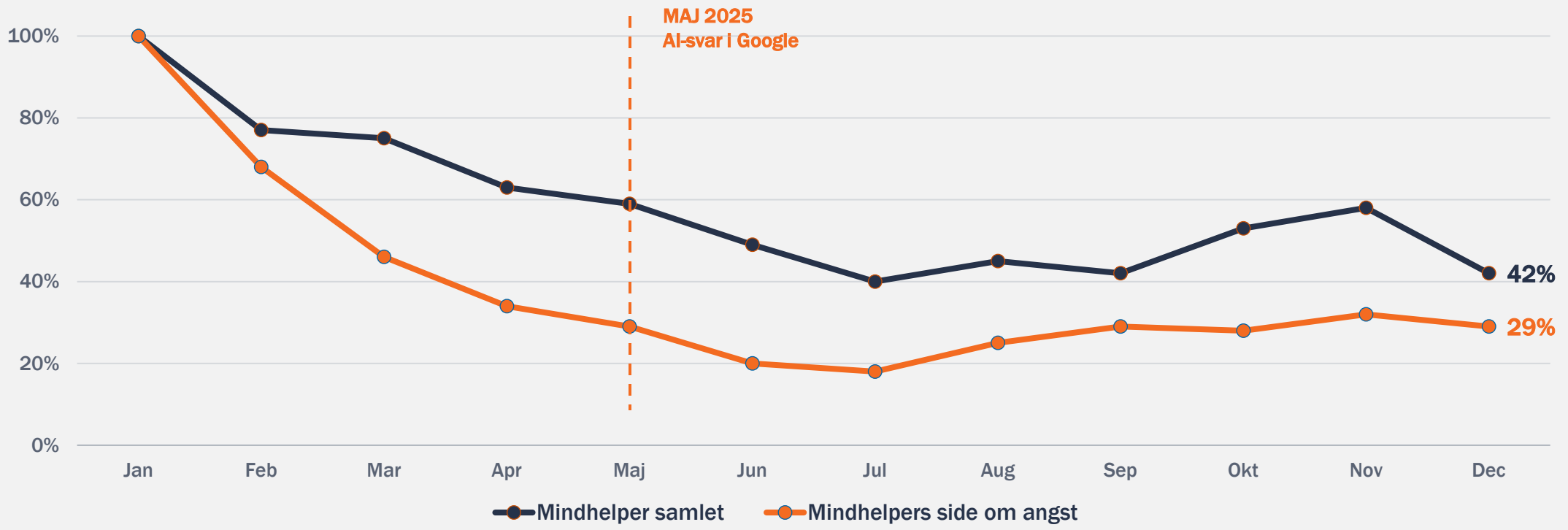
Center for Digital Psykiatri (2026). Mental sundhed: Sådan ændrer AI danskernes søgen efter viden og hjælp.

- **Anledning: Googles AI-svar i Danmark (maj 2025)**
- **Først intern strategiproces, men blev til publikation**
- **Kortlægger forandringerne i brugeradfærd**
- **4 mulige fremtidsscenarier og analyse på tværs**
- **Udgivet februar 2026, så meget er sket siden da, men analysen holder overordnet set stadig**

Center for Digital Psykiatri

2025 ændrede mindhelper.dks rolle i samfundet

Andel af trafikken i januar 2025



Center for Digital Psykiatri

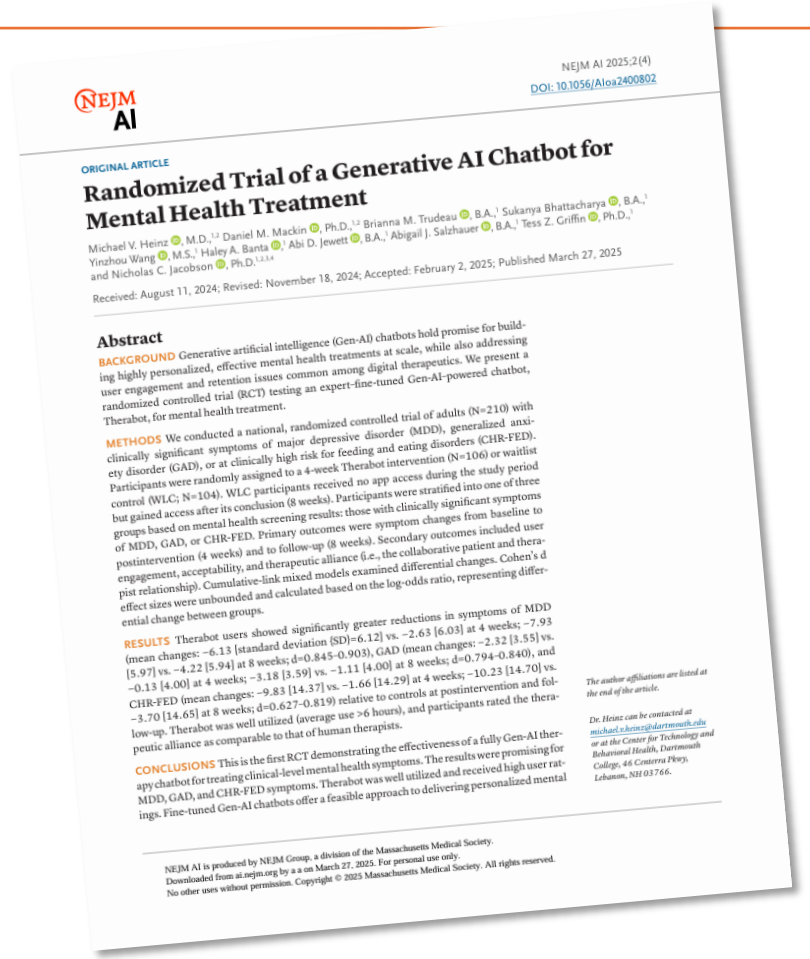
Andel af læste ord om mental sundhed, som er genereret af AI

3 ud af 4

Tors gæt-stimat på baggrund af:
Cloudflare, SE Rankings audit af Google AI Overviews, Agarwal & Sen (2026)

Center for Digital Psykiatri

Mulighed: Specialiseret chatbot + kliniker



NEJM AI 2025;2(4)
DOI: 10.1056/AIoa2400802

ORIGINAL ARTICLE

Randomized Trial of a Generative AI Chatbot for Mental Health Treatment

Michael V. Heinz, M.D.,[1,2] Daniel M. Mackin, Ph.D.,[1,2] Brianna M. Trudeau, B.A.,[1] Sukanya Bhattacharya, B.A.,[1] Yinzhou Wang, M.S.,[1] Haley A. Banta,[1] Abi D. Jewett, B.A.,[1] Abigail J. Salzhauer, B.A.,[1] Tess Z. Griffin, Ph.D.,[1] and Nicholas C. Jacobson, Ph.D.[1,2,3,4]

Received: August 11, 2024; Revised: November 18, 2024; Accepted: February 2, 2025; Published March 27, 2025

Abstract

BACKGROUND Generative artificial intelligence (Gen-AI) chatbots hold promise for building highly personalized, effective mental health treatments at scale, while also addressing user engagement and retention issues common among digital therapeutics. We present a randomized controlled trial (RCT) testing an expert-fine-tuned Gen-AI-powered chatbot, Therabot, for mental health treatment.

METHODS We conducted a national, randomized controlled trial of adults (N=210) with clinically significant symptoms of major depressive disorder (MDD), generalized anxiety disorder (GAD), or at clinically high risk for feeding and eating disorders (CHR-FED). Participants were randomly assigned to a 4-week Therabot intervention (N=106) or waitlist control (WLC; N=104). WLC participants received no app access during the study period but gained access after its conclusion (8 weeks). Participants were stratified into one of three groups based on mental health screening results: those with clinically significant symptoms of MDD, GAD, or CHR-FED. Primary outcomes were symptom changes from baseline to postintervention (4 weeks) and to follow-up (8 weeks). Secondary outcomes included user engagement, acceptability, and therapeutic alliance (i.e., the collaborative patient and therapist relationship). Cumulative-link mixed models examined differential changes. Cohen's d effect sizes were unbounded and calculated based on the log-odds ratio, representing differential change between groups.

RESULTS Therabot users showed significantly greater reductions in symptoms of MDD (mean changes: −6.13 [standard deviation (SD)=6.12] vs. −2.63 [6.03] at 4 weeks; −7.93 [5.97] vs. −4.22 [5.94] at 8 weeks; d=0.845–0.903), GAD (mean changes: −2.32 [3.55] vs. −0.13 [4.00] at 4 weeks; −3.18 [3.59] vs. −1.11 [4.00] at 8 weeks; d=0.794–0.840), and CHR-FED (mean changes: −9.83 [14.37] vs. −1.66 [14.29] at 4 weeks; −10.23 [14.70] vs. −3.70 [14.65] at 8 weeks; d=0.627–0.819) relative to controls at postintervention and follow-up. Therabot was well utilized (average use >6 hours), and participants rated the therapeutic alliance as comparable to that of human therapists.

CONCLUSIONS This is the first RCT demonstrating the effectiveness of a fully Gen-AI therapy chatbot for treating clinical-level mental health symptoms. The results were promising for MDD, GAD, and CHR-FED symptoms. Therabot was well utilized and received high user ratings. Fine-tuned Gen-AI chatbots offer a feasible approach to delivering personalized mental

The author affiliations are listed at the end of the article.

Dr. Heinz can be contacted at michael.v.heinz@dartmouth.edu or at the Center for Technology and Behavioral Health, Dartmouth College, 46 Centerra Pkwy, Lebanon, NH 03766.

NEJM AI is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from ai.nejm.org by a a on March 27, 2025. For personal use only.
No other uses without permission. Copyright © 2025 Massachusetts Medical Society. All rights reserved.

Heinz et. al. (2025)

- RCT med 210 voksne gennemført i foråret 2024
- Ikke generel chatbot: eksperttuning i manualiseret kognitiv adfærdsterapi og human-in-the-loop
- Store symptomfald mod venteliste efter 4 og 8 uger
- 95 % brugte løsningen med høj aktivitet og terapeutisk alliance på niveau med samtaleterapi
- Alle svar overvåget: 28 indgreb – få, men afgørende

Risiko: Fejl leveres i et overbevisende svar



Open access Original research

BMJ Open

Generative artificial intelligence-driven chatbots and medical misinformation: an accuracy, referencing and readability audit

Nicholas B Tiller ,[1] Alessandro R Marcon ,[2] Marco Zenone,[3] Kristin E Kidd,[4] Asker E Jeukendrup,[5] Zubin Master,[4,6,7,8] Timothy Caulfield[2]

To cite: Tiller NB, Marcon AR, Zenone M, *et al*. Generative artificial intelligence-driven chatbots and medical misinformation: an accuracy, referencing and readability audit. *BMJ Open* 2026;**16**:e112695. doi:10.1136/bmjopen-2025-112695

► Prepublication history and additional supplemental material for this article are available online. To view these files visit the journal online (https://doi.org/10.1136/bmjopen-2025-112695).

ZM and TC are joint senior authors.

Received 21 October 2025
Accepted 11 February 2026

Check for updates

© Author(s) (or their employer(s)) 2026. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

Correspondence to
Dr Nicholas B Tiller;
nicholas.tiller@lundquist.org

ABSTRACT

Objectives Artificial intelligence (AI)-driven chatbots have been rapidly adopted across research, education, business, marketing and medicine. Most interactions, however, come from non-experts using chatbots like search engines, including for everyday health and medical queries.

Design We conducted an original study to audit chatbot responses in health and medical fields prone to misinformation.

Methods Five popular chatbots were assessed: Gemini (Google), DeepSeek (High-Flyer), Meta AI (Meta), ChatGPT (OpenAI) and Grok (xAI). In February 2025, each chatbot was prompted with 10 questions from five categories: cancer, vaccines, stem cells, nutrition and athletic performance. We deployed an adversarial-like framework, using open- and closed-ended prompts designed to strain models toward misinformation or contraindicated advice. Two experts from each category rated responses as 'non-problematic', 'somewhat problematic' or 'highly problematic' using a coding matrix based on objective, predefined criteria. Citations were scored for accuracy and completeness, and each response was given a Flesch Reading Ease score.

Results Nearly half (49.6%) of responses were problematic: 30% somewhat problematic and 19.6% highly problematic. Response quality did not differ significantly among chatbots (p=0.566) but Grok generated significantly more highly problematic responses than would be expected under a random distribution (z-score +2.07, p=0.038). Performance was strongest in vaccines (mean z-score −2.57) and cancer (−2.12), and weakest in stem cells (+1.25), athletic performance (+3.74) and nutrition (+4.35). Chatbot outputs were consistently expressed with confidence and certainty; from 250 total questions, there were only two refusals to answer (0.8%), both from Meta AI. Reference quality was poor, with a median completeness score of 40% (Q1–Q3: 20–67%). Chatbot hallucinations and fabricated citations precluded any chatbot from producing a fully accurate reference list. All readability scores were graded as 'Difficult' (30–50), equivalent to college sophomore–senior level.

Conclusions The audited chatbots performed poorly when answering questions in misinformation-prone health and medical fields. Continued deployment without public education and oversight risks amplifying misinformation.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ The audit was broad, assessing responses from five publicly available AI chatbots across five misinformation-prone categories using two prompt types.
⇒ We developed a robust coding matrix to assess response accuracy using a three-tier 'any-failure' rating system that was highly sensitive to misleading content, thereby prioritising safety over precision.
⇒ Generative AI is evolving rapidly, and the chatbots assessed here reflect the models available at the time of the audit.
⇒ We prompted chatbots to return 'scientific references', which may have excluded legitimate sources of health information, such as technical reports, policy briefs, or publications from reputable medical institutions.

INTRODUCTION

Generative artificial intelligence (AI) has evolved from rudimentary algorithms designed for playing checkers into sophisticated large language models (LLMs) capable of generating original text, images, audio and code. Owing to their natural language comprehension and human-like responses, AI chatbots that are powered by LLMs have been rapidly adopted across sectors. Estimates are that three in four people use such chatbots in the workplace.[1] Students use them to complete assignments,[2] researchers to draft manuscripts,[3] and professionals to support decision-making in business and marketing. In medicine, AI chatbots have the potential to revolutionise healthcare delivery, especially in low- and middle-income countries,[4] and are increasingly used to support clinicians with documentation, decision-making and education.[5] AI chatbots can already surpass human experts in forecasting experimental outcomes,[6] making

BMJ Open: first published as 10.1136/bmjopen-2025-112695 on 14 April 2026. Downloaded from http://bmjopen.bmj.com/ on April 27, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

- 49,6 % af AI-svar var problematiske ifølge eksperter
- 19,6% af dem var meget problematiske
- Chatbots afviser nærmest aldrig at besvare et sundhedsspørgsmål
- Ingen chatbot leverede et eneste svar med fuldt korrekte kildehenvisninger

Tiller et. al. (2026)

Center for Digital Psykiatri

Risiko: Forstærker uhensigtsmæssig adfærd



nature medicine

Article

https://doi.org/10.1038/s41591-026-04577-2

A clinically validated framework for auditing AI chatbot behavior in mental health interactions

Received: 9 March 2026

Accepted: 9 July 2026

Published online: 07 August 2026

Check for updates

Veith Weilnhammer[1,2], Kevin YC Hou[3], Lennart Luettgau[3], Christopher Summerfield[3,4], Raymond Dolan[1] & Matthew M. Nour[1,5,6]

Millions of users turn to consumer artificial intelligence chatbots to discuss emotional, behavioral and mental-health concerns, creating an urgent need for rigorous and scalable safety evaluations. Here we introduce simulated (SIM) vulnerability-amplifying interaction loops (VAILs) (SIM-VAIL), a clinically validated framework for auditing chatbot behavior in mental-health contexts. SIM-VAIL simulates users with specific psychiatric vulnerabilities and conversational intents, engages them in multi-turn conversations with frontier artificial intelligence chatbots (including Claude, ChatGPT, Gemini, Grok and Llama models) and scores each exchange across 13 clinically grounded risk dimensions. Across 810 conversations, spanning 9 target chatbots and 30 simulated user profiles, concerning behavior in target chatbots was widespread, albeit reduced in newer models. Concerning behavior varied by user vulnerability and conversational intent, accumulated over turns, and could be reduced by interventions at early escalation points. Risk was highest when otherwise supportive chatbot behaviors reinforced the psychological mechanisms underlying the simulated user's vulnerability, a pattern we term a VAIL. SIM-VAIL provides a scalable framework for mapping mental-health risk across users, chatbots and conversational trajectories, offering a foundation for targeted safety improvements.

Access to mental-health care is severely limited[1]. With a global median of 13 mental-health workers per 100,000 people[2], existing support systems underserve people living with mental illness[3] and a far larger population seeking support for everyday behavioral health challenges[4]. Against this backdrop, millions now turn to general-purpose consumer artificial intelligence (AI) chatbots, such as ChatGPT, Claude, Gemini and Copilot, for emotional support, relationship guidance, companionship and therapeutic advice[4,5].

The widespread adoption, continuous availability and low marginal cost of consumer AI chatbots have raised hopes that they might supplement existing mental-health services or provide support where professional care is inaccessible[6–8]. These same properties, however, have been associated with mental-health risks to vulnerable users[9–11]. Consequently, academics, clinicians and industry actors increasingly recognize an urgent need for improved tools to evaluate and improve chatbot behavior in mental-health contexts[12–15].

AI chatbots are built on large language models (LLMs)—transformer-based deep neural networks whose outputs are probabilistic and context-dependent, making it impossible to guarantee in advance how a model will behave in a new situation[16]. Empirical safety

[1]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London, UK. [2]Sydney Medical School, University of Sydney, Sydney, Australia. [3]UK AI Security Institute, London, UK. [4]Department of Experimental Psychology, University of Oxford, Oxford, UK. [5]Microsoft AI, London, UK. [6]Department of Psychiatry, University of Oxford, Oxford, UK. e-mail: v.weilnhammer@ucl.ac.uk; matthew.nour@psych.ox.ac.uk

- Vi har alle haft det på fornemmelsen, nu er det dokumenteret...
- Chatbots forstærker uhensigtsmæssige tanke- og handlemønstre med et "Vulnerability-Amplifying Interaction Loop"
- Mønsteret kan ikke læses af ét svar, men bliver tydeligt, imens chatbotten tilpasser sig brugeren
- 810 samtaler med 9 populære chatbots via 30 simulerede brugerprofiler

Weilnhammer et. al. (2026)

Center for Digital Psykiatri

Nej, de bliver ikke nødvendigvis bedre...



MindBench.ai

August 2026

HealthBench-Psych: A Mental Health Subset of OpenAI's HealthBench

Matthew Flathers[1] Phuong Anh Nguyen[1] Jill Noorily[1] Julian Herpertz[1]
Meiting Chen[1] Jasreen Multani[1] Samuel Powell[2]
Mason Granof[1] Mark Kalinch[1] John Torous[1]
[1]Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA
[2]Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA

arXiv:2608.25071v1 [cs.CL] 25 Aug 2026

Abstract

General-purpose health benchmarks increasingly anchor claims about LLM medical performance, but they are not always resolved by clinical specialty, making domain-specific performance hard to isolate. Mental health is of acute public-health concern as millions of people turn to LLMs for psychological support, and most existing evaluations are bespoke academic benchmarks that are difficult to integrate into developer workflows. We introduce HealthBench-Psych and HealthBench-Psych-Hard. We screened HealthBench's 5,000 physician-rubric conversations for mental-health relevance with a transparent LLM-applied rubric, then validated the subset through two rounds of blinded clinician review with concealed known-exclude controls, yielding 610 conversations (12.2% of the corpus). Evaluating 20 frontier and open models under a cross-vendor panel of three LLM judges, we find a statistically tied frontier cluster, measurable refusal behavior in two models, and near-identical rankings across judges ($r \geq 0.92$). We release the subset, pipeline, model responses, grades, and analysis code as a reusable resource.

1 Introduction

People increasingly bring mental-health concerns to general-purpose LLMs, often in moments of acute distress and outside any clinical pathway (Bodner et al., 2026; Rousmaniere et al., 2026). Evaluating how these systems behave in mental-health conversations is therefore pressing. Yet the existing mental-health evaluation literature clusters narrowly on suicide-risk detection and diagnostic classification (Flathers et al., 2026), and risk classification and board-style diagnostic reasoning are poor proxies for the open-ended, ambiguous, multi-turn help-seeking that likely constitutes most real usage (Raji et al., 2025; Stade et al., 2024; Luo et al., 2025).

Conversational health benchmarks have begun to close this realism gap. HealthBench (Arora et al., 2025) is the most prominent example, comprising 5,000 multi-turn conversations, each response graded against physician-authored rubrics. Its format is free-text rather than saturated multiple-choice (Singhal et al., 2025), and the benchmark is openly licensed. But it carries no specialty metadata. Its item tags describe conversational behavior (hedging, emergency referral, context-seeking) and physician-uncertainty structure, never specific medical topic. This matters because HealthBench has begun to appear in AI company model cards and technical reports as the core of their health testing, making it one of developers' most widely used mental-health evaluation instruments (OpenAI, 2026; Anthropic, 2026; Meta Superintelligence Labs, 2026). A lab reporting a strong HealthBench number may therefore appear competent in psychiatry and clinical psychology, with no clear way to isolate that domain performance on its own.

We address this gap in two stages. We introduce HealthBench-Psych, a clinician-adjudicated subset of 610 mental-health-relevant HealthBench conversations, and use it for the first specialty-resolved comparison of 20 frontier and open models under a cross-vendor panel of three LLM judges. We contribute: (1) a transparent, reproducible pipeline for

Preprint v1.0.0.
Code: https://github.com/mindbench-ai/healthbench-psych
Data: https://huggingface.co/datasets/mindbench-ai/healthbench-psych
Correspondence: mflather@bidmc.harvard.edu

Flathers et. al. (2026)

- Nyt preprint, der sammenligner 20 sprogmodeller på et helt nyt benchmark for mental sundhed
- Mental sundhed: Claude Fable 5 < Claude Opus 5. Kimi K3 < Kimi K2.6. GPT-5.6-sol < GPT-5.5
- I vanskelige samtaler om mental sundhed falder kvaliteten af svar helt ned til omkring 0,40 (0,0-1,0)
- Kun 13 ud af 12.200 gange afviste en chatbot at svare på et spørgsmål om mental sundhed

Center for Digital Psykiatri

Anbefalinger

1. Forstå hvor meget adfærden omkring mental sundhed er ændret på 1,5 år.
2. Nye sundhedstilbud skal matche den nye brugeradfærd og teknologi
3. Det offentlige skal have få brugerflader og én ”digital fordør til sundhedsvæsenet”