

Paolo Raile

KI in der Psychotherapie: Ersatz, Werkzeug oder Projektionsfläche?



Psychotherapie-Wissenschaft

16. Jahrgang, Nr. 2, 2026, Seite 7–15

DOI: 10.30820/1664-9583-2026-2-7

Psychosozial-Verlag

Impressum

Psychotherapie-Wissenschaft

eISSN 1664-9591

16. Jahrgang Heft 2/2026

<https://doi.org/10.30820/1664-9583-2026-2>

info@psychotherapie-wissenschaft.info

www.psychotherapie-wissenschaft.info

Herausgeber

Herausgeberverein der Zeitschrift

Psychotherapie-Wissenschaft

Peter Schulthess

Zanggerweg 33

CH-8006 Zürich

Mobile + 41 76 559 19 20

E-Mail: peter.schulthess@psychotherapie-wissenschaft.info

Redaktion

Mara Foppoli, Lugano

Lea-Sophie Richter, Zürich

Mario Schlegel, Zürich

Peter Schulthess, Zürich

Nataliya Tereshchenko, Zürich/Wien

Hinweise für AutorInnen befinden
sich auf der Homepage der Zeitschrift:
www.psychotherapie-wissenschaft.info

Verlag

Psychosozial-Verlag GmbH & Co. KG

Walltorstr. 10

D-35390 Gießen

+49 6421 96 99 78 26

info@psychosozial-verlag.de

www.psychosozial-verlag.de

Anzeigen

Anfragen zu Anzeigen bitte an den Verlag:

anzeigen@psychosozial-verlag.de

Es gelten die Preise der auf www.psychosozial-verlag.de
einsehbaren Mediadaten.

Gönnermitglieder des Herausgebervereins

wenden sich bitte direkt an die Redaktion:

redaktion@psychotherapie-wissenschaft.info

Datenbanken

Die Zeitschrift Psychotherapie-Wissenschaft wird regelmäßig
in der Abstract-Datenbank PsycInfo der American Psycho-
logical Association (APA), im Directory of Open Journals
(DOAJ) und in den Publikationsdatenbanken PSYNDEX und
PsychArchives des Leibniz-Institut für Psychologie/Leibniz
Institute for Psychology (ZPID) erfasst.



Die Beiträge dieser Zeitschrift sind unter der Creative Commons
Attribution-NonCommercial-NoDerivs 3.0 DE Lizenz lizenziert.
Diese Lizenz erlaubt die private Nutzung und unveränderte
Weitergabe, verbietet jedoch die Bearbeitung und kommerzielle
Nutzung. Weitere Informationen finden Sie unter:
creativecommons.org/licenses/by-nc-nd/3.0/de

KI in der Psychotherapie: Ersatz, Werkzeug oder Projektionsfläche?

Technische Grundlagen, ethische Implikationen, klinische Perspektiven

Paolo Raile

Psychotherapie-Wissenschaft 16 (2) 2026 7–15

www.psychotherapie-wissenschaft.info

CC BY-NC-ND

<https://doi.org/10.30820/1664-9583-2026-2-7>

Zusammenfassung: Large Language Models (LLMs) werden als aktuell besonders sichtbarer Bereich generativer KI zunehmend in der psychotherapeutischen Versorgung eingesetzt, etwa für Screening, Triage, Psychoedukation, dokumentationsbezogene Assistenz sowie zur Simulation therapeutischer Interaktionen. Systematische Übersichten zeigen, dass LLMs in klar strukturierten, sprachbasierten Aufgaben teils Leistungen erbringen, die mit denen klinischer Fachkräfte vergleichbar sind, insbesondere in der Diagnoseunterstützung bei Depression und PTSD sowie in der Strukturierung klinischer Fallkonzeptionen. Zugleich berichten Nutzer:innen über eine hohe wahrgenommene Zugänglichkeit, Niedrigschwelligkeit und empathische Gesprächsführung. Demgegenüber stehen zentrale Limitationen: Halluzinationen, inkonsistente Antwortqualität, Bias entlang sozialer Kategorien, fehlende Transparenz der Trainingsdaten, ungeklärte Fragen von Datenschutz und Haftung sowie eine strukturelle Grenze hinsichtlich affektiver Empathie, Intentionalität und klinischer Verantwortungsfähigkeit. In Krisensituationen – etwa bei Suizidalität oder akuter Psychose – zeigen LLMs teils inadäquate oder riskante Empfehlungen. Der aktuelle wissenschaftliche Konsens bewertet LLMs daher als potenziell sinnvolle assistive Werkzeuge unter fachlicher Supervision, nicht jedoch als Ersatz für die therapeutische Beziehung oder für komplexe, autonome Behandlung. Regulatorische und berufsethische Leitplanken, insbesondere Transparenz, Datenschutz, menschliche Aufsicht und klare Verantwortungszuordnung, sind für eine sichere Integration zentral. Insgesamt eröffnen LLMs neue Möglichkeiten für Screening, Psychoedukation, Dokumentation und Ausbildung, während klinische Sicherheit, ethische Rahmenbedingungen und die Rolle der therapeutischen Beziehung weiterhin kritisch reflektiert und empirisch untersucht werden müssen.

Schlüsselwörter: KI, Psychotherapie, Ethik, Large Language Models

Künstliche Intelligenz (KI) hat in den vergangenen Jahren erheblich an gesellschaftlicher und fachlicher Relevanz gewonnen, nicht zuletzt durch die rasche Verbreitung generativer Systeme. In der Allgemeinbevölkerung zeigt sich dabei ein ambivalentes, überwiegend positives Bild: KI wird breit genutzt und vielfach als hilfreich wahrgenommen, insbesondere dort, wo Effizienzgewinne oder Qualitätsverbesserungen erwartet werden (Gesk & Leyer, 2022; Jo, 2023; Kelly et al., 2022; Liehner et al., 2023; Vo et al., 2023; Wu et al., 2023). Gleichzeitig bestehen deutliche Vorbehalte, v. a. in sensiblen Kontexten wie Medizin, Behörden oder Arbeitswelt (Xiang et al., 2020; Yigitcanlar et al., 2023). Systematische Reviews zeigen, dass Bürger:innen im Gesundheitsbereich klare Präferenzen für Modelle der «geteilten Verantwortung» äussern, also für eine Kombination aus KI-Unterstützung und menschlicher Expertise plädieren, während rein KI-basierte Entscheidungen skeptischer gesehen werden (Chen et al., 2022; Vo et al., 2023; Young et al., 2021; Beets et al., 2023; Wu et al., 2023). Als zentrale Einflussfaktoren gelten wahrgenommener Nutzen und Vertrauen in die verantwortlichen Institutionen, während Datenschutzsorgen, Angst vor Fehlentscheidungen, Entmenschlichung und Arbeitsplatzverlust die Akzeptanz mindern (Catalina et al., 2023; Ismatullaev & Kim, 2022;

Kelly et al., 2022; Liehner et al., 2023; Rana et al., 2024; Tam et al., 2025; Wu et al., 2023; Young et al., 2021). Gleichzeitig fühlen sich viele Menschen fachlich unsicher, was KI eigentlich ist und wie sie funktioniert, obwohl sie diese im Alltag bereits regelmässig – oft unbewusst – nutzen (Kelly et al., 2022; Koenig, 2024; Liehner et al., 2023). Insgesamt lässt sich somit von einer vorsichtig-positiven, aber nicht unkritischen Grundhaltung sprechen, die stark vom Kontext abhängt.

Auch unter Psychotherapeut:innen, Psycholog:innen und Psychiatr:innen findet sich ein vergleichbar differenziertes Bild. Studien zeigen eine grundsätzlich vorsichtig-positive Einschätzung, insbesondere hinsichtlich potenzieller Unterstützung bei Diagnostik, Behandlungsplanung, Monitoring, Dokumentation und administrativer Entlastung (Alimour et al., 2024; Cecil et al., 2025; Kleine et al., 2023; Linardon et al., 2025; Stroud et al., 2025; Wagner & Schwind, 2025). Gleichzeitig bestehen deutliche Wissens- und Vertrauenslücken sowie z. T. ausgeprägte Vorbehalte, etwa in Bezug auf Fehlriskiken, Haftung, Intransparenz («Black Box»), Datenschutz und potenziellen Missbrauch (Cecil et al., 2025; Heinrichs et al., 2025; Rogan et al., 2024; Stroud et al., 2025). Ein breiter fachlicher Konsens verweist darauf, dass KI nicht als Ersatz für

die therapeutische Beziehung verstanden werden sollte, da diese weiterhin als zentraler Wirkfaktor gilt (Cecil et al., 2025; Linardon et al., 2025; Rogan et al., 2024; Stroud et al., 2025; Wagner & Schwind, 2025). Unterschiede in der Akzeptanz entstehen u. a. durch Technikaffinität, wahrgenommenen Nutzen sowie Berufsgruppe – Psychiater:innen zeigen tendenziell höhere Nutzungsintentionen als Psychotherapeut:innen und klinische Psycholog:innen (Alimour et al., 2024; Blease et al., 2024; Cecil et al., 2025). Zugleich weisen Blindtests darauf hin, dass KI-generierte Beratungstexte – sofern deren Herkunft nicht bekannt ist – mitunter ähnlich empathisch oder qualitativ wie Expert:innentexte wahrgenommen werden können, was sowohl Potenzial als auch Klärungsbedarf verdeutlicht (Föylen et al., 2025). Insgesamt scheint sich der Gedanke durchzusetzen, dass KI eher als Werkzeug oder «Co-Pilot» denn als eigenständige Behandlerinstanz verstanden werden sollte.

Vor diesem Hintergrund zeigt sich in der Fachcommunity zunehmend eine Polarisierung: Während einige Kolleg:innen KI als potenziell entlastendes und unterstützendes Werkzeug einordnen, äussern andere deutliche Skepsis bis hin zu Ersetzungsängsten und Identitätsbedenken. Ein nicht zu unterschätzender Aspekt ist dabei, dass selbst viele fachlich interessierte Therapeut:innen nur begrenztes Wissen über die technischen Grundlagen moderner KI-Systeme besitzen. Studien zu Psychologiestudierenden und angehenden Psychotherapeut:innen weisen ebenfalls darauf hin, dass Akzeptanz und Nutzungsbereitschaft eng mit Wissen, wahrgenommener Nützlichkeit und konkreter Erfahrung zusammenhängen (Gado et al., 2021; Kleine et al., 2023). Diese Wissenslücken können zu unrealistischen Erwartungen, Verunsicherung oder pauschaler Ablehnung beitragen und erschweren eine sachliche, fachlich fundierte Auseinandersetzung (Cecil et al., 2025; Kelly et al., 2022; Liehner et al., 2023; Wagner & Schwind, 2025). Ziel dieses Beitrags ist es daher, eine theorie- und literaturbasierte Einordnung der Rolle von KI in der Psychotherapie zu bieten. Im Zentrum stehen drei Leitfragen: 1. Inwieweit können KI-Systeme menschliche Therapeutinnen – theoretisch oder praktisch – ersetzen, und wo liegen strukturelle Grenzen? 2. In welchen Bereichen kann KI bei verantwortungsvoller Nutzung für Therapeut:innen wie Patient:innen einen realistischen Mehrwert bieten? 3. Wie kann ein verständlicher Überblick über die technischen Grundlagen dazu beitragen, KI zu entmystifizieren und Polarisierungstendenzen zu reduzieren? KI wird dabei weder als Bedrohung noch als Heilsversprechen verstanden, sondern als Werkzeug, dessen Potenzial und Risiken einer reflektierten, fachlich und ethisch fundierten Diskussion bedürfen.

Wie funktionieren LLMs?

Bevor wir uns speziell Large Language Models (LLMs) anschauen, lohnt ein kurzer Blick auf das grössere Feld: KI umfasst alle rechnerbasierten Verfahren, die aus Daten lernen, Muster erkennen oder Entscheidungen unterstützen. Man kann KI auf einer ersten Ebene nach dem Lern-

prinzip unterscheiden. Dazu gehören z. B. überwachtes Lernen (das System lernt aus Beispielen mit bekannten Antworten), unüberwachtes Lernen (das System entdeckt Muster ohne vorgegebene Antworten) und Reinforcement Learning, bei dem ein System über Belohnung und Rückmeldung lernt (Janiesch et al., 2021; Yang et al., 2021; Çelik & Eltawil, 2024; Pternea et al., 2024). Innerhalb dieser Lernformen gibt es grob zwei grosse Gruppen von Modellen: diskriminative Modelle, die v. a. Dinge unterscheiden oder vorhersagen (z. B. «Depression ja/nein»), und generative Modelle, die neue Inhalte erzeugen können – also das, was heute meist unter «Generative KI» verstanden wird (Bandi et al., 2023; Harshvardhan et al., 2020; Yazdani et al., 2025). Zu diesen generativen Modellen gehören verschiedene Familien: etwa Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs) oder Diffusionsmodelle, die Bild- und Videoinhalte Schritt für Schritt aus Rauschen rekonstruieren (Bandi et al., 2023; Yazdani et al., 2025; Çelik & Eltawil, 2024). LLMs gehören ebenfalls in diese Gruppe: Sie sind generative Modelle für Sprache, die Token für Token neue Texte erzeugen und häufig auf der sog. Transformer-Architektur basieren (Banh & Strobel, 2023; Deldjoo et al., 2024; Raiaan et al., 2024). Diese Einordnung zeigt: LLMs sind ein spezifischer Zweig im vielschichtigen Feld der KI, und generative KI ist wiederum eine Teilmenge des gesamten KI-Spektrums.

Neben LLMs umfasst das weitere Feld KI-gestützter Anwendungen in Psychotherapie und Mental Health auch prädiktive Modelle, digitale Phänotypisierung, passive Sensorik, Sprach-, Bild- und Videoanalyse, klinische Entscheidungsunterstützung sowie Systeme zur Verlaufserfassung und Risikoeinschätzung. Diese Verfahren unterscheiden sich deutlich in Datenbasis, Zweck, Transparenz und klinischer Verantwortung. LLMs stehen im Folgenden im Vordergrund, weil sie derzeit eine besonders sichtbare Schnittstelle zwischen generativer KI, Sprache und therapeutischer Kommunikation bilden und damit zentrale Fragen nach Beziehung, Verantwortung, Projektion und klinischer Sicherheit bündeln.

Ganz basal betrachtet ist ein LLM eine sehr grosse Datei voller Zahlen, die während des Trainings aus unzähligen Textbeispielen gelernt hat, welche Wörter oder Wortteile (Tokens) typischerweise aufeinander folgen. Das Modell speichert keine Texte im herkömmlichen Sinn, sondern lernt mathematische Strukturen und Zusammenhänge in Sprache, sodass es später bei einer Eingabe Schritt für Schritt das wahrscheinlich nächste Token vorhersagen kann (Li et al., 2024; Lysyi, 2025; Rhee, 2025). Diese Datei kann – ausreichend Rechenleistung vorausgesetzt – auch lokal ohne Internetverbindung ausgeführt werden, denn das Gelernte steckt vollständig in ihren Parametern. Wenn man dem Modell also eine Frage stellt, wird diese zuerst in Tokens zerlegt, mathematisch dargestellt und anschliessend berechnet das System mit jedem Schritt, welches Token jetzt am wahrscheinlichsten ist. Man kann sich ein LLM daher als extrem dicht gepackten, statistischen Sprachraum vorstellen: kein Nachschlagewerk, sondern ein System, das sprachliche Wahrscheinlichkeit berechnet.

Stellen wir uns zunächst vor, Sie geben einem KI-System die Frage ein: «Wer war Sigmund Freud?» Was passiert jetzt im Inneren? Ein LLM macht etwas scheinbar Einfaches: Es versucht Schritt für Schritt vorherzusagen, welches Wort bzw. welcher Wortteil wahrscheinlich als Nächstes kommt. Das Modell hat zuvor aus sehr vielen Textbeispielen gelernt, in welchen sprachlichen Zusammenhängen welche Wörter typischerweise auftreten. Technisch nennt man das autoregressive Next-Token-Vorhersage (ebd.). Tokens sind dabei kleine Spracheinheiten – oft Wörter oder Wortteile. Auch unsere Frage wird zuerst in solche Tokens zerlegt. Diese Tokens werden nicht als Wörter verarbeitet, sondern in Zahlen umgewandelt, sog. Vektoren. Zusätzlich wird für jedes Token gespeichert, an welcher Position es in der Eingabe steht, z.B. ob es am Anfang oder am Ende der Frage auftaucht (Zheng et al., 2024; Kashyap, 2025). Das Modell «weiss» also mathematisch, welche Begriffe in welcher Reihenfolge stehen, ohne diese menschlich zu verstehen. Auf dieser Grundlage beginnt es nun, die nächste sprachliche Einheit vorherzusagen. Es berechnet etwa, wie wahrscheinlich es ist, dass nach «Sigmund Freud war ...» die Tokens «ein», «österreichischer», «Neurologe» oder «Psychiater» folgen – basierend auf gelernten Mustern. Wichtig ist: Das Modell sagt nicht die Wahrheit voraus, sondern sprachliche Wahrscheinlichkeit. Das ist der Kern dessen, was ein LLM tut (Li et al., 2024; Lysyi, 2025; Rhee, 2025).

Der entscheidende Mechanismus dafür heisst Selbst-Attention. Er lässt sich so verstehen: Wenn das Modell überlegt, welches Wort jetzt sinnvoll wäre, «schaut» jedes Token auf die anderen Tokens im bisherigen Text und bewertet, welche davon besonders wichtig sind. Mathematisch geschieht das über drei Vektorarten – Query, Key und Value –, die jeweils bestimmte Rollen in dieser Kontextbewertung spielen (Neo et al., 2024; Zhang et al., 2025; Smaldone et al., 2025). Das klingt abstrakt, ist aber gut mit unserem Beispiel zu erklären. Steht in der Eingabe «Sigmund Freud», wird das Modell frühere Vorkommen dieses Namens in den Trainingsdaten «erinnern» – nicht als konkrete Stellen, sondern als statistische Muster – und diese Information in die nächste Vorhersage einfließen lassen. Es wird daher Tokens wie «Psychoanalyse», «Begründer», «Psychiater» oder «Österreich» mit hoher Wahrscheinlichkeit in Betracht ziehen. Dabei gilt eine wichtige Regel: Das Modell darf nur auf bereits erschienene Tokens achten, nicht auf zukünftige – es kann also nicht in die Zukunft «spicken». Dies nennt man kausale Maskierung (Barbero et al., 2025; Li et al., 2024). Moderne Modelle verwenden viele parallele «Aufmerksamkeitsköpfe». Man kann sich das so vorstellen, dass jeder Kopf auf unterschiedliche sprachliche Aspekte schaut – einer vielleicht auf Grammatik, ein anderer auf Namen, ein weiterer auf thematische Zusammenhänge (Neo et al., 2024; Zheng et al., 2024). Aus diesen parallelen Bewertungen entsteht schliesslich eine kontextsensitive Repräsentation, die dem Modell hilft, die nächste sprachliche Einheit auszuwählen. Forschung zeigt, dass bestimmte Attention-Köpfe besonders stark mit der Aktivierung jener Modellteile verbunden sind, die

die nächste Tokenvorhersage direkt beeinflussen (Neo et al., 2024). So entsteht Schritt für Schritt eine Antwort, die menschlich wirkt – in unserem Beispiel vielleicht: «Sigmund Freud war ein österreichischer Neurologe und Begründer der Psychoanalyse.»

Am Ende jeder dieser Rechenschritte steht eine Wahrscheinlichkeitsverteilung über alle möglichen nächsten Tokens, die mittels Softmax berechnet wird (Lysyi, 2025). Doch nun muss das Modell sich entscheiden. Es kann entweder immer das wahrscheinlichste Token wählen – dann klingt die Antwort sehr stabil, aber manchmal etwas eintönig. Oder es kann durch sog. Sampling-Verfahren (z.B. Top-k- oder Top-p-Sampling) ein wenig Zufall zulassen, sodass alternative, aber weiterhin plausible Formulierungen entstehen. Über die Temperatur wird eingestellt, wie stark dieser Zufall ist: Je höher die Temperatur, desto «kreativer», aber auch unvorhersehbarer wird die Antwort (ebd.). Dieser Prozess wiederholt sich so lange, bis der Satz oder Text beendet ist. Allerdings ist das Verfahren rechenaufwendig, weil jedes Token auf viele andere Tokens achten muss – weshalb verschiedene effizientere Varianten von Attention entwickelt wurden, etwa lineare oder «sparse» Attention-Verfahren, die denselben Grundmechanismus nutzen, aber weniger Rechenleistung benötigen (Sun et al., 2025; Zhang et al., 2025; Kashyap, 2025). Wichtig bleibt jedoch: Das Modell arbeitet immer mit Wahrscheinlichkeiten, nie mit Bedeutungen im menschlichen Sinn. Wenn es also über Freud schreibt, dann deshalb, weil es gelernt hat, dass Menschen in Texten Freud typischerweise so beschreiben – nicht, weil es «weiss», wer Freud war.

Für die psychotherapeutische Praxis bedeutet dieses Funktionsprinzip etwas sehr Zentrales: Ein LLM rekonstruiert sprachliche Muster. Es kann daher Antworten erzeugen, die informativ, flüssig und kontextbezogen klingen. Gleichzeitig erklärt dieses Prinzip auch typische Fehler wie sog. Halluzinationen: Wenn es keine passenden Muster gibt oder Wahrscheinlichkeitsverteilungen unscharf sind, kann das Modell plausible, aber sachlich falsche Aussagen erzeugen. Die zugrunde liegende Mechanik bleibt dabei dieselbe – Sprache wird auf Basis mathematischer Muster vorhergesagt (Li et al., 2024; Lysyi, 2025; Rhee, 2025). Die Beziehungsgestaltung, Intentionalität oder subjektives Erleben, die menschliche Psychotherapie auszeichnen, sind in solchen Systemen jedoch nicht vorhanden. Gerade deshalb ist es wichtig, die Funktionsweise nüchtern zu verstehen: Sie entmystifiziert KI und ermöglicht es, das Werkzeug kompetent, kritisch und verantwortungsvoll einzusetzen.

Das Beispiel «Wer war Sigmund Freud?» lässt sich daher als didaktische Zusammenfassung lesen: Die Eingabe wird tokenisiert, in numerische Repräsentationen übersetzt, mit Positionsinformationen versehen und anschliessend Token für Token fortgeführt. In jedem Schritt entsteht eine Wahrscheinlichkeitsverteilung über mögliche Fortsetzungen; das System wählt – deterministisch oder über Sampling – die sprachlich plausibelste nächste Einheit aus (ebd.). Für die klinische Einordnung ist entscheidend, dass diese Plausibilität nicht mit mensch-

lichem Wissen, Verstehen oder Verantwortungsfähigkeit verwechselt wird. Gerade daraus ergeben sich Potenziale für strukturierte sprachliche Unterstützung, aber auch Risiken wie Halluzinationen, Scheinsicherheit und fehlende Einsicht in eigene Grenzen.

LLMs in der Psychotherapie

LLMs haben in den vergangenen Jahren Eingang in vielfältige Bereiche der psychotherapeutischen Versorgung gefunden. Sie werden eingesetzt zur Unterstützung diagnostischer Prozesse, für Screening- und Triage-Entscheidungen, für Psychoedukation, für niedrigintensive Interventionen in manualisierten Verfahren, als Assistenzsysteme für Therapeut:innen sowie zur Simulation therapeutischer Gespräche in Aus- und Weiterbildung. Systematische Übersichtsarbeiten zeigen, dass LLMs in sprachbasierten Aufgaben teilweise Leistungen erbringen, die mit klinischen Fachkräften vergleichbar sind, insbesondere in klar strukturierten und textnahen Settings (Guo et al., 2024; Hua et al., 2024, 2025; Jin et al., 2025). In Vignettenstudien können LLMs bei bestimmten Störungen, etwa depressiven Episoden oder posttraumatischen Belastungsstörungen, diagnostische Genauigkeit auf oder nahe dem Niveau menschlicher Behandler:innen erreichen und plausible, guideline-kompatible Behandlungsoptionen vorschlagen (Gabriel et al., 2024; Levkovich, 2025; Omar et al., 2024). Auf dieser Ebene scheinen LLMs besonders dort zu profitieren, wo diagnostische Informationen bereits sprachlich explizit vorliegen, Kategorien klar definiert sind und wenig komplexe, non-verbale oder implizite Beziehungssignale interpretiert werden müssen. Auch in der Zusammenfassung grosser Textmengen – etwa elektronischer Krankenakten, Beratungsprotokollen oder Sitzungsnotizen – sowie in der Strukturierung klinischer Argumentationen zeigen LLMs ein hohes Leistungsniveau (Adhikary et al., 2024; Hua et al., 2024; Jin et al., 2025). Feingetunte Modelle wie Mental-Alpaca, Mental-FLAN-T5 oder PsyLLM, die explizit auf psychotherapeutische Dialogdaten trainiert wurden, scheinen dabei generischen Modellen überlegen zu sein und erreichen in Benchmarks Beratungsqualität, die sich teilweise ChatGPT-ähnlichem Niveau annähert (Hu et al., 2025; Liu et al., 2023; Xu et al., 2023).

Eine weitere Stärke von LLMs liegt im Bereich der sprachlichen Interaktion und niedrigschwelligen Zugänglichkeit. Nutzer:innen erleben LLM-basierte Chatbots meist als leicht erreichbar, anonym, nicht verurteilend und 24/7 verfügbar – Faktoren, die insbesondere bei stigmatisierten oder psychosozial belasteten Menschen die Kontaktaufnahme erleichtern können (Guo et al., 2024; Jung et al., 2025). Viele Betroffene berichten in qualitativen Studien, dass sie LLMs zur emotionalen Entlastung, zur Vorbereitung von Therapiesitzungen oder zum Einüben von Skills – etwa aus der kognitiven Verhaltenstherapie – nutzen (Ma et al., 2023; Song et al., 2024). Auch evidenzbasierte therapeutische Mikroskills wie kognitive Empathie, Validierung oder problemorientierte Strukturie-

rung lassen sich in LLM-Antworten beobachten. Mehrere Studien kommen zu dem Befund, dass LLM-Antworten in schriftlicher Kommunikation hinsichtlich wahrgenommener Empathie in einzelnen Szenarien sogar höher bewertet werden als Antworten menschlicher Fachkräfte, sofern die Urheberschaft nicht offengelegt wird (Föyen et al., 2025; Gabriel et al., 2024; Sorin et al., 2023). Diese Ergebnisse legen nahe, dass LLMs kognitive Empathie – also das Erkennen und sprachliche Spiegeln emotionaler Zustände – in hohem Mass simulieren können. Gleichzeitig betonen ethische Analysen die Differenz zwischen simulierter und gelebter Empathie: LLMs verfügen weder über eigene Intentionalität noch über affektives Erleben; Beziehung wird ausschliesslich durch die Modellstatistik erzeugt (Rudra et al., 2025; Orrú et al., 2025).

Neben diesen Interaktions- und Screeningfunktionen gewinnen LLMs auch als Assistenzsysteme für Psychotherapeut:innen an Bedeutung. Sie können klinische Informationen bündeln, Falldarstellungen strukturieren, psychoedukative Materialien erstellen, Behandlungsoptionen skizzieren oder Mikroskills in Therapiesitzungen annotieren (Hammerfald et al., 2026; Volkmer et al., 2024). Einige Reviews argumentieren, dass LLMs damit zur Entlastung in überlasteten Versorgungssystemen beitragen könnten, etwa indem Routineaufgaben wie Dokumentation, Termin-Vorbereitungen oder standardisierte Psychoedukation teilweise automatisiert werden (Guo et al., 2024; Lawrence et al., 2024; Malgaroli et al., 2025). Besonders in Ausbildung und Supervision können LLM-basierte Simulationen von Patient:innen und Therapeut:innen hilfreich sein, um Gesprächsführung zu üben und differenziertes Feedback zu erhalten (Haider et al., 2025; Na et al., 2025). Raile (2024) konnte in einer hermeneutischen Analyse realer Interaktionen zwischen ChatGPT und Psychotherapeut:innen bzw. Patient:innen zeigen, dass ChatGPT in vielen Fällen als ergänzendes Werkzeug wahrgenommen wird – sowohl von Behandler:innen, die es zur Reflexion klinischer Fälle nutzen, als auch von Patient:innen, die es zwischen Sitzungen als stützende Ressource einsetzen.

Gleichzeitig besteht ein breiter wissenschaftlicher Konsens, dass diese Potenziale von erheblichen klinischen und ethischen Grenzen begleitet werden. Zunächst handelt es sich bei LLMs strukturell um statistische Modelle der nächsten Token-Wahrscheinlichkeit. Sie verfügen weder über Bewusstsein noch über gelebte Emotionalität, Intentionalität oder Verantwortungsfähigkeit (Orrú et al., 2025). Sie können daher zwar plausible Therapiegespräche führen, tragen jedoch keine klinische Verantwortung im Sinne menschlicher Entscheidungskompetenz. Zudem zeigen Studien, dass LLMs bei komplexen, frühen oder atypischen Störungsbildern deutlich schlechtere diagnostische Leistungen erbringen als in klar umrissenen Fällen. Für prodromale Psychosen wurden diagnostische Genauigkeiten um etwa 55 % berichtet – ein Wert, der für klinische Entscheidungen unzureichend ist (Levkovich, 2025). Hinzu kommt eine teilweise erhebliche Inkonstanz: Dasselbe Modell erzeugt für ähnliche Eingaben mitunter variierende oder widersprüchliche Antworten, die jedoch

mit hoher sprachlicher Sicherheit formuliert werden. Dieses Phänomen der «halluzinierten» Sicherheit erhöht das Risiko, dass Nutzer:innen unzutreffende Aussagen für verlässlich halten (Guo et al., 2024; Jung et al., 2025; Mündler et al., 2023).

Ein zweiter Problembereich betrifft Bias und Ungleichheit. LLMs reproduzieren – abhängig von Trainingsdaten und Modellarchitektur – gesellschaftliche Vorurteile, etwa entlang von «Race», Gender oder Stigma psychischer Erkrankungen. Es existieren Hinweise darauf, dass LLMs ungleiche Empathielevel gegenüber unterschiedlichen Gruppen zeigen und stereotype Deutungsmuster verstärken können (Gabriel et al., 2024; Lawrence et al., 2024). Dadurch besteht die Gefahr, bestehende Versorgungsungleichheiten zu verfestigen, insbesondere wenn LLM-basierte Systeme vorrangig in wohlhabenden und technologisch entwickelten Gesundheitssystemen entstehen (Margaroli et al., 2025). Auch Datenschutz, Intransparenz der Trainingsdaten und unklare Haftungsfragen werden in nahezu allen Übersichtsarbeiten als zentrale Barrieren benannt (Guo et al., 2024; Jin et al., 2025).

Die therapeutische Beziehung markiert einen weiteren Kernpunkt der Debatte. Die therapeutische Allianz zählt über Verfahren hinweg zu den robustesten Wirkfaktoren; Meta-Analysen zeigen einen konsistenten Zusammenhang zwischen Allianzqualität und Behandlungsergebnis (Flückiger et al., 2018). Während LLMs kognitive Empathie überzeugend simulieren, bleibt affektive Beteiligung im Sinne miterlebter Betroffenheit unerreichbar. Mehrere Arbeiten argumentieren, dass authentische therapeutische Haltung – verstanden als verantwortliches Sich-Einlassen mit biografischer, leiblicher und sozialer Verankerung – strukturell nicht abbildbar ist (Rudra et al., 2025; Orrú et al., 2025). Patient:innen erleben LLM-Unterstützung in Studien durchaus als hilfreich, zugleich aber häufig als emotional «leer», selbst wenn die Antworten formal empathisch sind (Rudra et al., 2025). Ethik-Autor:innen empfehlen daher, LLM-Empathie primär als «passive» Empathie – also das Erkennen, Markieren und Eskalieren emotionaler Zustände – zu gestalten und die Illusion einer genuine Beziehung möglichst zu vermeiden (ebd.). Zusätzlich ist zu beachten, dass therapeutische Entwicklung häufig nicht nur von stabiler Allianz, sondern auch von der Bearbeitung von Allianzrupturen lebt; solche Spannungs- und Reparaturprozesse setzen Responsivität, Verantwortungsübernahme und Beziehungsgedächtnis voraus (Safran et al., 2011).

Besonders kritisch werden Krisensituationen und akute Störungsbilder eingeschätzt. Mehrere Studien zeigen, dass generische Therapie-Chatbots bei Suizidalität, Selbstverletzung oder psychotischen Symptomen unangemessene oder potenziell gefährliche Empfehlungen ausgeben können (Lawrence et al., 2024; Ohu et al., 2025). Aus Sicht der Forschungslandschaft gilt daher als Konsens: Für autonome Behandlung komplexer oder akuter psychischer Erkrankungen fehlt derzeit jede belastbare Evidenz, und die Risiken überwiegen den Nutzen deutlich (Guo et al., 2024; Hua et al., 2025; Omar et al., 2024). Die meisten Autor:innen sehen LLMs entsprechend eher als

Assistenz- und Ergänzungswerkzeuge unter fachlicher Supervision – etwa für Screening, Monitoring, Psychoedukation, Dokumentation oder Ausbildung – und nicht als Ersatz für menschliche Therapeut:innen (Lawrence et al., 2024; Na et al., 2025).

Interessant ist in diesem Kontext die Frage nach der Akzeptanz durch Fachkräfte. Befragungsstudien unter Psychotherapeut:innen, Psycholog:innen und Psychiater:innen zeigen überwiegend eine vorsichtig-positive Grundhaltung gegenüber KI als Werkzeug, bei gleichzeitiger Skepsis gegenüber einem Ersatz der menschlichen Beziehung (Wagner & Schwind, 2025; Cecil et al., 2025; Stroud et al., 2025). Besonders positiv werden diagnostische Unterstützung, Risikovorhersagen, Dokumentationsentlastung und Praxisorganisation bewertet, während Datenschutz, Black-Box-Charakter, Haftungsfragen und «AI-Angst» zu Vorbehalten beitragen (Cecil et al., 2025; Rogan et al., 2024). Die Technikaffinität der Behandler:innen erweist sich als wesentlicher Prädiktor für Akzeptanz, während psychodynamisch orientierte Psychotherapeut:innen tendenziell skeptischer bleiben als Psychiater:innen (Cecil et al., 2025; Alimour et al., 2024). Diese Ambivalenz spiegelt sich auch in der wissenschaftlichen Literatur wider: LLMs werden als nützliches Werkzeug anerkannt, gleichzeitig aber klar als kein adäquater Ersatz für die spezifische Qualität menschlicher therapeutischer Beziehung verstanden (Wagner & Schwind, 2025; Lawrence et al., 2024).

LLMs können im psychotherapeutischen Feld einen kognitiv starken, empathisch klingenden, strukturierten Ko-Therapie-Charakter einnehmen – insbesondere in schriftlich-sprachlichen, klar definierten Aufgabenbereichen. Sie erweitern den Zugang zu Unterstützung, erleichtern Screening- und Dokumentationsprozesse und bieten realistische Trainingssimulationen. Ihre strukturellen Begrenzungen – fehlende Intentionalität, Halluzinationen, Bias, mangelnde Verantwortbarkeit und die Unersetzbarkeit gelebter Beziehung – markieren jedoch eine ethische und klinische Grenze. Vor diesem Hintergrund empfehlen nahezu alle Reviews eine assistive, supervidierte Nutzung von LLMs in der psychotherapeutischen Versorgung und warnen vor einer vorschnellen Substitution menschlicher Behandlung durch autonome KI-Systeme (Guo et al., 2024; Hua et al., 2025; Lawrence et al., 2024).

Orientierung für eine verantwortungsvolle Integration von LLMs in die Psychotherapie

Die vorliegenden Entwicklungen im Bereich der LLMs markieren zweifellos einen technologischen Paradigmenwechsel, der auch die psychotherapeutische Versorgung nachhaltig beeinflusst. Die dargestellten Befunde zeigen, dass LLMs in klar strukturierten, sprachbasierten Anwendungen ein hohes Leistungsniveau erreichen können. Sie unterstützen Screening- und Triageprozesse, strukturieren klinische Informationen, generieren psychoedukative Inhalte und sind in der Lage, manualisierte Gesprächsanteile – insbesondere aus kognitiv-verhaltenstherapeutischen Traditionen – in sprachlich kohärenter Weise

zu reproduzieren (Guo et al., 2024; Hua et al., 2024, 2025; Jin et al., 2025; Lawrence et al., 2024). Gleichzeitig wird deutlich, dass Nutzer:innen LLMs häufig als niedrigschwellig, verfügbar und nicht-wertend erleben und diese Eigenschaften den Zugang zu Unterstützung teilweise erleichtern (Guo et al., 2024; Jung et al., 2025). Diese Befunde begründen weder Alarmismus noch Heilsversprechen, sondern eine differenzierte Einschätzung: LLMs sind leistungsstarke sprachliche Werkzeuge, die in bestimmten Anwendungsfeldern einen realen Versorgungsbeitrag leisten können.

Gleichzeitig verdeutlichen die Ergebnisse die strukturellen Grenzen dieser Systeme. Die Funktionsweise der Modelle beruht auf statistischer Wahrscheinlichkeitsvorhersage und nicht auf verstehender, intentionaler oder affektiv eingebetteter Beziehungsgestaltung. Damit ist erklärbar, warum LLMs auch in komplexen klinischen Kontexten plausible und empathisch klingende Sprache erzeugen können, ohne über Verantwortungsfähigkeit, Gewissen oder erlebte Beteiligung zu verfügen (Orrú et al., 2025; Rudra et al., 2025). Diese grundlegende Differenz ist klinisch bedeutsam: Als sprachlich-kognitive Simulation können LLMs therapeutische Haltungen nachahmen; sie können sie jedoch nicht verkörpern. Dies wird besonders dort relevant, wo Beziehung, Affektregulation, Responsivität und das Aushalten von Nicht-Wissen zentrale Wirkfaktoren darstellen. Auch das Phänomen der «halluzinierten Sicherheit» zeigt, dass LLM-Antworten formal kompetent erscheinen können, ohne epistemisch zuverlässig zu sein – ein Umstand, der insbesondere in vulnerablen Patient:innengruppen Risiken birgt (Guo et al., 2024; Jung et al., 2025; Mündler et al., 2023). In Krisensituationen – etwa bei Suizidalität oder akuter Psychose – sind die Risiken unzureichender, inkonsistenter oder nicht verantwortbarer Antworten deutlich dokumentiert, weshalb eine autonome klinische Nutzung derzeit nicht vertretbar erscheint (Lawrence et al., 2024; Ohu et al., 2025; Hua et al., 2025; Omar et al., 2024).

Vor diesem Hintergrund zeichnet sich in der Literatur eine dreifache Funktionseinteilung ab. *Erstens* der Bereich assistiver Nutzung, in dem LLMs Verantwortung nicht übernehmen, sondern unterstützen. Hierzu zählen diagnostische Strukturierung, Screening, Monitoring, Zusammenfassung klinischer Dokumentation, administrative Entlastung, Psychoedukation und Simulation in der Aus- und Weiterbildung (Guo et al., 2024; Lawrence et al., 2024; Malgaroli et al., 2025; Na et al., 2025). Dieser Bereich wird in der Regel positiv bewertet und ist – unter Beachtung klarer datenschutzrechtlicher, organisatorischer und ethischer Rahmenbedingungen – als klinisch anschlussfähig zu verstehen. *Zweitens* existiert der Bereich ko-therapeutischer Nutzung, etwa in manualisierten Interventionen oder als begleitendes Tool in Langzeitbehandlungen. Hier weisen mehrere Autor:innen darauf hin, dass eine solche Nutzung nur unter professioneller Supervision, mit klarer Verantwortungszuordnung und definierter Indikationsstellung sinnvoll ist (Lawrence et al., 2024; Na et al., 2025). *Drittens* besteht der Bereich autonomer Behandlung – also der Einsatz ohne mensch-

liche Verantwortungsinstanz. Die vorliegende Evidenz spricht deutlich dafür, dass dieser Bereich derzeit weder wissenschaftlich abgesichert noch ethisch vertretbar ist, insbesondere bei Krisen, komplexer Komorbidität oder schwerer Psychopathologie (Guo et al., 2024; Hua et al., 2025; Lawrence et al., 2024).

Eine besondere Rolle nimmt die Frage der therapeutischen Beziehung ein. Mehrere Studien weisen darauf hin, dass LLM-Antworten als empathisch, validierend und unterstützend wahrgenommen werden können (Föyen et al., 2025; Gabriel et al., 2024; Sorin et al., 2023). Gleichwohl betont die ethische Literatur die Differenz zwischen simulierter Empathie und affektiv miterlebter Beziehung (Rudra et al., 2025). Aus klinischer Perspektive dürfte ein professionell verantworteter Umgang darin bestehen, diese Differenz nicht zu verschleiern, sondern transparent zu halten: LLMs können empathische Marker setzen, dürfen jedoch nicht als beziehungsfähige Subjekte erscheinen. In ähnlicher Weise ist die Frage nach Bias, Gerechtigkeit und Ungleichheit zentral. Da LLMs aus bestehenden Datenstrukturen lernen, reproduzieren sie auch deren Verzerrungen. Es gibt Hinweise darauf, dass Antworten je nach sozialer Kategorie systematisch variieren können (Gabriel et al., 2024; Lawrence et al., 2024). Insofern besteht die Gefahr einer Reproduktion struktureller Ungleichheiten innerhalb psychischer Versorgung, wenn KI-Systeme unreflektiert implementiert werden (Malgaroli et al., 2025).

Aus psychodynamischer Perspektive wird KI damit auch zur Projektionsfläche. Gerade weil LLMs sprachlich responsiv, jederzeit verfügbar und scheinbar nicht wertend reagieren, können ihnen Verstehen, Neutralität, Fürsorge, Autorität oder sogar eine quasi-intentionale Haltung zugeschrieben werden. Solche Zuschreibungen sind klinisch nicht per se problematisch; sie können entlastend wirken, Reflexion anstossen oder Übergangsräume für Selbstexploration eröffnen. Zugleich bergen sie Risiken, wenn die simulierte Responsivität als echte Beziehung missverstanden wird oder wenn Übertragungs- und Idealisierungsprozesse unbemerkt bleiben. Für psychodynamische und beziehungsorientierte Forschung ergibt sich daraus eine wichtige Fragestellung: Nicht nur die technische Leistungsfähigkeit von KI ist relevant, sondern auch, welche inneren Objekte, Beziehungserwartungen und Affektregulationsmuster in der Interaktion mit KI aktiviert, stabilisiert oder verändert werden.

Parallel dazu dokumentiert die Literatur die Einstellungen professioneller Akteur:innen. Psychotherapeut:innen, Psycholog:innen und Psychiater:innen beschreiben KI weitgehend als potenziell hilfreiches Werkzeug, weisen aber auf erhebliche Unsicherheiten, Wissenslücken und ethische Bedenken hin. Sie betonen die Unersetzbarkeit der therapeutischen Beziehung und schätzen KI v. a. dort, wo sie diagnostische, organisatorische und administrative Prozesse unterstützt (Cecil et al., 2025; Stroud et al., 2025; Wagner & Schwind, 2025). Einstellungen variieren dabei entlang von Berufsgruppe, Technikaffinität und wahrgenommenem Nutzen (Alimour et al., 2024; Cecil et al., 2025). Diese Befunde unterstreichen die Relevanz einer aufgeklärten, entmystifizierten Auseinandersetzung mit

KI, in der weder Überhöhung noch pauschale Ablehnung dominieren.

Aus diesen empirischen Befunden und ethischen Erwägungen lassen sich klinische Mindeststandards formulieren. *Erstens* sollte jede Nutzung von LLMs transparent sein – Patient:innen müssen wissen, wann KI-basierte Systeme eingesetzt werden. *Zweitens* bleibt die klinische Verantwortung unteilbar menschlich: KI kann Entscheidungsprozesse unterstützen, nicht aber Verantwortlichkeit tragen. *Drittens* sind klare Notfall- und Eskalationsmechanismen erforderlich. *Viertens* müssen Datenschutz und Sicherheit höchsten Standards genügen, insbesondere im Hinblick auf sensible Gesundheitsdaten; hierzu gehören die Vorgaben der DSGVO/GDPR, nationale Datenschutzgesetze wie das DSG sowie berufsrechtliche Schweige- und Sorgfaltspflichten. *Fünftens* ist kontinuierliche Qualitäts- und Risikoevaluation notwendig, da Modelle sich technisch wie inhaltlich verändern können. Diese Orientierung entspricht auch internationalen und europäischen Leitlinien, die für KI im Gesundheitswesen Transparenz, menschliche Aufsicht, Nicht-Schädigung, Fairness, Datenschutz und Rechenschaftspflicht betonen (European Parliament and Council of the European Union, 2016, 2024; WHO, 2021). *Sechstens* sollte eine Überidentifikation mit KI-Antworten aktiv vermieden werden – sowohl auf Patient:innen- als auch auf Therapeut:innen-Seite.

Aus wissenschaftlicher Perspektive ergeben sich mehrere Forschungslücken. Es fehlen valide Wirksamkeitsstudien zu LLM-basierten Interventionen, insbesondere im Vergleich zu evidenzbasierten Therapien, sowie Daten zur Langzeitwirksamkeit und Schadensprävention. Auch Prozessforschung – etwa zu Übertragung, Gegenübertragung, Bindungsschemata oder verinnerlichten Relationen in KI-gestützten Gesprächen – steckt noch in den Anfängen. Darüber hinaus sind bias-kritische Analysen, klinisch-ethische Frameworks und rechtliche Klärungen dringend erforderlich (Guo et al., 2024; Hua et al., 2025; Lawrence et al., 2024; Malgaroli et al., 2025).

Insgesamt zeichnet sich als klinisch-ethische Orientierung ein klarer Rahmen ab: LLMs können die psychotherapeutische Versorgung sinnvoll ergänzen, aber sie verändern nicht den Kern psychotherapeutischer Arbeit. Dieser Kern besteht weiterhin in verantwortlicher menschlicher Beziehungsgestaltung, geteiltem Bedeutungsverstehen, ethischer Haltung und professioneller Entscheidungskompetenz. LLMs können dazu beitragen, Zeitressourcen freizusetzen, Zugänge zu erleichtern, Informationen zu strukturieren und niedrigschwellige Unterstützung bereitzustellen. Sie sind jedoch weder beziehungs- noch verantwortungsfähig und damit strukturell nicht geeignet, die therapeutische Dyade zu ersetzen. Eine assistive, kritisch-reflektierte Nutzung unter klinischer Supervision scheint deshalb gegenwärtig der angemessene Weg zu sein – verbunden mit der Verpflichtung, die technologischen Entwicklungen aufmerksam, wissenschaftlich fundiert und ethisch verantwortlich zu begleiten (Guo et al., 2024; Hua et al., 2025; Lawrence et al., 2024; Wagner & Schwind, 2025).

Literatur

- Adhikary, P.K. et al. (2024). Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study. *JMIR Mental Health*, 11, e57306. <https://doi.org/10.2196/57306>
- Alimour, S. et al. (2024). The quality traits of artificial intelligence operations in predicting mental healthcare professionals' perceptions: A case study in the psychotherapy division. *Journal of Autonomous Intelligence*. <https://doi.org/10.32629/jai.v7i4.1438>
- Bandi, A. et al. (2023). The power of generative AI: A review of requirements, models, input-output formats, evaluation metrics, and challenges. *Future Internet*, 15, 260. <https://doi.org/10.3390/fi15080260>
- Banh, L. & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33, 1–17. <https://doi.org/10.1007/s12525-023-00680-1>
- Barbero, F. et al. (2025). Why do LLMs attend to the first token? *arXiv*. <https://doi.org/10.48550/arxiv.2504.02732>
- Beets, B. et al. (2023). Surveying public perceptions of artificial intelligence in health care in the United States: Systematic review. *Journal of Medical Internet Research*, 25. <https://doi.org/10.2196/40337>
- Blease, C. et al. (2024). Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: An online mixed methods survey. *Psychiatry Research*, 333. <https://doi.org/10.1016/j.psychres.2024.115724>
- Catalina, Q. et al. (2023). Knowledge and perception of the use of AI and its implementation in the field of radiology: Cross-sectional study. *Journal of Medical Internet Research*, 25. <https://doi.org/10.2196/50728>
- Cecil, J. et al. (2025). Mental health practitioners' perceptions and adoption intentions of AI-enabled technologies: An international mixed-methods study. *BMC Health Services Research*, 25. <https://doi.org/10.1186/s12913-025-12715-8>
- Çelik, A. & Eltawil, A. (2024). At the dawn of generative AI era: A tutorial-cum-survey on new frontiers in 6G wireless intelligence. *IEEE Open Journal of the Communications Society*, 5, 2433–2489. <https://doi.org/10.1109/ojcoms.2024.3362271>
- Chen, M. et al. (2022). Acceptance of clinical artificial intelligence among physicians and medical students: A systematic review with cross-sectional survey. *Frontiers in Medicine*, 9. <https://doi.org/10.3389/fmed.2022.990604>
- Deldjoo, Y. et al. (2024). Recommendation with generative models. *arXiv*. <https://doi.org/10.48550/arxiv.2409.15173>
- European Parliament and Council of the EU (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. *Official Journal of the EU*.
- European Parliament and Council of the EU (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the EU*.
- Flückiger, C. et al. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, 55(4), 316–340. <https://doi.org/10.1037/pst0000172>
- Föyén, L. et al. (2025). Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. *Internet Interventions*, 41. <https://doi.org/10.1016/j.invent.2025.100841>
- Gabriel, S. et al. (2024). Can AI relate? Testing large language model response for mental health support. *arXiv*. <https://doi.org/10.48550/arxiv.2405.12021>
- Gado, S. et al. (2021). Artificial intelligence in psychology: How can we enable psychology students to accept and use artificial intelligence? *Psychology Learning & Teaching*, 21, 37–56. <https://doi.org/10.1177/14757257211037149>
- Gesk, T. & Leyer, M. (2022). Artificial intelligence in public services: When and why citizens accept its usage. *Government Information Quarterly*, 39, 101704. <https://doi.org/10.1016/j.giq.2022.101704>

- Guo, Z. et al. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*, 11. <https://doi.org/10.2196/57400>
- Hammerfald, K. et al. (2026). Leveraging large language models to identify microcounseling skills in psychotherapy transcripts. *Psychotherapy Research*, 36(6), 1058–1076. <https://doi.org/10.1080/10503307.2025.2539405>
- Haider, S. et al. (2025). Synthetic patient–physician conversations simulated by large language models: A multi-dimensional evaluation. *Sensors*, 25. <https://doi.org/10.3390/s25144305>
- Harshvardhan, G. et al. (2020). A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38, 100285. <https://doi.org/10.1016/j.cosrev.2020.100285>
- Heinrichs, H. et al. (2025). Physicians' attitudes toward artificial intelligence in medicine: Mixed methods survey and interview study. *Journal of Medical Internet Research*, 27. <https://doi.org/10.2196/74187>
- Hu, H. et al. (2025). Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv*. <https://doi.org/10.48550/arxiv.2505.15715>
- Hua, Y. et al. (2024). Large language models in mental health care: A scoping review. *Current Treatment Options in Psychiatry*, 12. <https://doi.org/10.1007/s40501-025-00363-y>
- Hua, Y. et al. (2025). A scoping review of large language models for generative tasks in mental health care. *NPJ Digital Medicine*, 8. <https://doi.org/10.1038/s41746-025-01611-4>
- Ismatullaev, U. & Kim, S. (2022). Review of the factors affecting acceptance of AI-infused systems. *Human Factors*, 66, 126–144. <https://doi.org/10.1177/00187208211064707>
- Janiesch, C. et al. (2021). Machine learning and deep learning. *Electronic Markets*, 31, 685–695. <https://doi.org/10.1007/s12525-021-00475-2>
- Jin, Y. et al. (2025). The applications of large language models in mental health: Scoping review. *Journal of Medical Internet Research*, 27, e69284. <https://doi.org/10.2196/69284>
- Jo, H. (2023). Understanding AI tool engagement: A study of ChatGPT usage and word-of-mouth among university students and office workers. *Telematics and Informatics*, 85, 102067. <https://doi.org/10.1016/j.tele.2023.102067>
- Jung, K. et al. (2025). «I've talked to ChatGPT about my issues last night»: Examining mental health conversations with large language models through Reddit analysis. *PACM HCI*, 9. <https://doi.org/10.1145/3757537>
- Kashyap, A. (2025). Recurrent memory-augmented transformers with chunked attention for long-context language modeling. *arXiv*. <https://doi.org/10.48550/arxiv.2507.00453>
- Kelly, S. et al. (2022). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, 77, 101925. <https://doi.org/10.1016/j.tele.2022.101925>
- Kleine, A. et al. (2023). Attitudes toward the adoption of artificial intelligence-enabled mental health tools among prospective psychotherapists: Cross-sectional study. *JMIR Human Factors*, 10. <https://doi.org/10.2196/46859>
- Koenig, P. (2024). Attitudes toward artificial intelligence: Combining three theoretical perspectives on technology acceptance. *AI & Society*, 40, 1333–1345. <https://doi.org/10.1007/s00146-024-01987-z>
- Lawrence, H. et al. (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11. <https://doi.org/10.2196/59479>
- Levkovich, I. (2025). Evaluating diagnostic accuracy and treatment efficacy in mental health: A comparative analysis of large language model tools and mental health professionals. *European Journal of Investigation in Health, Psychology and Education*, 15. <https://doi.org/10.3390/ejihpe15010009>
- Li, Y. et al. (2024). Mechanics of next token prediction with self-attention. *arXiv*. <https://doi.org/10.48550/arxiv.2403.08081>
- Liehner, G. et al. (2023). Perceptions, attitudes and trust toward artificial intelligence. An assessment of the public opinion. *Artificial Intelligence and Social Computing*. <https://doi.org/10.54941/ahfe1003271>
- Linardon, J. et al. (2025). Current practices and perspectives of artificial intelligence in the clinical management of eating disorders: Insights from clinicians and community participants. *The International Journal of Eating Disorders*, 58, 724–734. <https://doi.org/10.1002/eat.24385>
- Liu, J. et al. (2023). ChatCounselor: A large language model for mental health support. *arXiv*. <https://doi.org/10.48550/arxiv.2309.15461>
- Lysyi, P. (2025). Principles of large language models (LLM). *Journal of Numerical and Applied Mathematics*. <https://doi.org/10.17721/2706-9699.2025.1.06>
- Ma, Z. et al. (2023). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *arXiv*. <https://doi.org/10.48550/arxiv.2307.15810>
- Malgaroli, M. et al. (2025). Large language models for the mental health community: Framework for translating code to care. *The Lancet Digital Health*, 7, e282–e285. [https://doi.org/10.1016/S2589-7500\(24\)00255-3](https://doi.org/10.1016/S2589-7500(24)00255-3)
- Mündler, N. et al. (2023). Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv*. <https://doi.org/10.48550/arxiv.2305.15852>
- Na, H. et al. (2025). A survey of large language models in psychotherapy: Current landscape and future directions. *Findings of the Association for Computational Linguistics: ACL 2025*, 7362–7376. <https://doi.org/10.18653/v1/2025.findings-acl.385>
- Neo, C. et al. (2024). Interpreting context look-ups in transformers: Investigating attention-MLP interactions. *arXiv*. <https://doi.org/10.48550/arxiv.2402.15055>
- Ohu, F.C. et al. (2025). Public health risk management, policy, and ethical imperatives in the use of AI tools for mental health therapy. *Healthcare*, 13(21), 2721. <https://doi.org/10.3390/healthcare13212721>
- Omar, M. et al. (2024). Applications of large language models in psychiatry: A systematic review. *Frontiers in Psychiatry*, 15, 1422807. <https://doi.org/10.3389/fpsy.2024.1422807>
- Orrú, G. et al. (2025). Large language models and psychiatry. *International Journal of Law and Psychiatry*, 101, 102086. <https://doi.org/10.1016/j.ijlp.2025.102086>
- Pternea, M. et al. (2024). The RL/LLM taxonomy tree: Reviewing synergies between reinforcement learning and large language models. *arXiv*. <https://doi.org/10.1613/jair.1.15960>
- Raiaan, M. et al. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12, 26839–26874. <https://doi.org/10.1109/access.2024.3365742>
- Raile, P. (2024). The usefulness of ChatGPT for psychotherapists and patients. *Humanities and Social Sciences Communications*, 11, 47. <https://doi.org/10.1057/s41599-023-02567-0>
- Rana, M. et al. (2024). Assessing AI adoption in developing country academia: A trust- and privacy-augmented UTAUT framework. *Heliyon*, 10. <https://doi.org/10.1016/j.heliyon.2024.e37569>
- Rhee, P. (2025). Moving beyond next-token prediction: Transformers are context-sensitive language generators. *arXiv*. <https://doi.org/10.48550/arxiv.2504.10845>
- Rogan, J. et al. (2024). Health care professionals' views on the use of passive sensing, AI, and machine learning in mental health care: Systematic review with meta-synthesis. *JMIR Mental Health*, 11. <https://doi.org/10.2196/49577>
- Rudra, P. et al. (2025). Large language models for surgical informed consent: An ethical perspective on simulated empathy. *Journal of Medical Ethics*. <https://doi.org/10.1136/jme-2024-110652>
- Safran, J.D. et al. (2011). Repairing alliance ruptures. *Psychotherapy*, 48(1), 80–87. <https://doi.org/10.1037/a0022140>
- Smaldone, A. et al. (2025). A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation. *Journal of Chemical Theory and Computation*. <https://doi.org/10.1021/acs.jctc.5c00331>
- Song, I. et al. (2024). The typing cure: Experiences with large language model chatbots for mental health support. *arXiv*. <https://doi.org/10.48550/arxiv.2401.14362>

- Sorin, V. et al. (2023). Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26. <https://doi.org/10.2196/52597>
- Stroud, A. et al. (2025). Physician perspectives on the potential benefits and risks of applying artificial intelligence in psychiatric medicine: Qualitative study. *JMIR Mental Health*, 12. <https://doi.org/10.2196/64414>
- Sun, Y. et al. (2025). Efficient attention mechanisms for large language models: A survey. *arXiv*. <https://doi.org/10.48550/arxiv.2507.19595>
- Tam, L. et al. (2025). Support for businesses' use of artificial intelligence: Dynamics of trust, distrust, and perceived benefits. *Media and Communication*. <https://doi.org/10.17645/mac.9534>
- Vo, V. et al. (2023). Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. *Social Science & Medicine*, 338, 116357. <https://doi.org/10.1016/j.socscimed.2023.116357>
- Volkmer, S. et al. (2024). Large language models in psychiatry: Opportunities and challenges. *Psychiatry Research*, 339, 116026. <https://doi.org/10.1016/j.psychres.2024.116026>
- Wagner, J. & Schwind, A. (2025). Investigating psychotherapists' attitudes towards artificial intelligence in psychotherapy. *BMC Psychology*, 13. <https://doi.org/10.1186/s40359-025-03071-7>
- WHO (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. <https://www.who.int/publications/i/item/9789240029200>
- Wu, C. et al. (2023). Public perceptions on the application of artificial intelligence in healthcare: A qualitative meta-synthesis. *BMJ Open*, 13. <https://doi.org/10.1136/bmjopen-2022-066322>
- Xiang, Y. et al. (2020). Implementation of artificial intelligence in medicine: Status analysis and development suggestions. *Artificial Intelligence in Medicine*, 102, 101780. <https://doi.org/10.1016/j.artmed.2019.101780>
- Xu, X. et al. (2023). Mental-LLM: Leveraging large language models for mental health prediction via online text data. *arXiv*. <https://doi.org/10.48550/arXiv.2307.14385>
- Yang, X. et al. (2021). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35, 8934–8954. <https://doi.org/10.1109/tkde.2022.3220219>
- Yazdani, S. et al. (2025). Generative AI in depth: A survey of recent advances, model variants, and real-world applications. *Journal of Big Data*, 12. <https://doi.org/10.1186/s40537-025-01247-x>
- Yigitcanlar, T. et al. (2023). Artificial intelligence in local government services: Public perceptions from Australia and Hong Kong. *Government Information Quarterly*, 40, 101833. <https://doi.org/10.1016/j.giq.2023.101833>
- Young, A. et al. (2021). Patient and general public attitudes towards clinical artificial intelligence: A mixed methods systematic review. *The Lancet Digital Health*, 3(9), e599–e611. [https://doi.org/10.1016/s2589-7500\(21\)00132-1](https://doi.org/10.1016/s2589-7500(21)00132-1)
- Zhang, S. et al. (2025). Curse of high dimensionality issue in transformer for long-context modeling. *arXiv*. <https://doi.org/10.48550/arxiv.2505.22107>
- Zheng, Z. et al. (2024). Attention heads of large language models. *Patterns*, 6. <https://doi.org/10.1016/j.patter.2025.101176>

AI in Psychotherapy:

Substitute, tool, or projection screen?

Technical fundamentals, ethical implications, clinical perspectives

Abstract: Large language models (LLMs) are increasingly being used in psychotherapeutic care, for example for screening, triage, psychoeducation, documentation-related assistance, and to simulate therapeutic interactions. Systematic reviews show that LLMs sometimes perform comparably to clinical professionals in clearly structured, language-based tasks, particularly in supporting the diagnosis of depression and PTSD and in structuring clinical case concepts. At the same time, users report high perceived accessibility, low threshold, and empathetic conversation skills. However, there are also key limitations: hallucinations, inconsistent response quality, bias along social categories, lack of transparency of training data, unresolved issues of data protection and liability, and a structural limitation with regard to affective empathy, intentionality, and clinical accountability. In crisis situations – such as suicidality or acute psychosis – LLMs sometimes provide inadequate or risky recommendations. The current scientific consensus therefore considers LLMs to be potentially useful assistive tools under professional supervision, but not as a substitute for the therapeutic relationship or for complex, autonomous treatment. Surveys of psychotherapists reveal a cautiously positive attitude toward LLMs as support tools, coupled with skepticism about substitution scenarios. Overall, LLMs open up new possibilities for screening, psychoeducation, documentation, and training, while clinical safety, ethical frameworks, and the role of the therapeutic relationship must continue to be critically reflected upon and empirically investigated.

Keywords: AI, psychotherapy, ethics, large language models

Biografische Notiz

Priv.-Doz. Dr. Dr. *Paolo Raile*, MSc., studierte Psychotherapiewissenschaft an der Sigmund-Freud-PrivatUniversität Wien (SFU), Soziale Arbeit an der Donau-Universität Krems und Europäische Ethnologie an der Universität Wien. Er forscht an der SFU, ist Autor wissenschaftlicher Texte, Psychotherapeut, Sozialarbeiter, Lebens- und Sozialberater sowie Gründer und Leiter zweier psychosozialer Organisationen in Wien.

Kontakt

E-Mail: paolo@raile.at; paolo.raile@sfu.ac.at