

Review

Konvergensi *Large Language Models* dan *Generative Artificial Intelligence* dalam Psikoterapi: Keamanan Klinis, Efektivitas Terapeutik, dan Dinamika Interaksi Manusia

Indri Wijayanti^{1*}, Annastasya Nabila Elsa Wulandari²

1 Program Studi Keperawatan Anestesiologi, Universitas Harapan Bangsa, Purwokerto, Indonesia

2 Program Studi Informatika, Universitas Harapan Bangsa, Purwokerto, Indonesia

* Correspondence: indriwijayanti@uhb.ac.id

Abstract: The integration of large language models (LLMs) and generative AI into mental health services is advancing faster than the process of consolidating clinical evidence and establishing an adequate governance framework. This article presents a narrative review that synthesizes empirical, technical, and conceptual literature on the convergence of LLMs with psychotherapy practice across four main domains: clinical effectiveness, safety and bias, human-AI relational dynamics, and architecture and model deployment. The literature indicates that LLMs possess strong communicative capabilities and are often perceived as having significant potential. However, this potential is not yet supported by an adequate clinical evidence base, as the literature remains dominated by observational designs, scenario-based evaluations, and short-term perception measures. Evidence from controlled clinical trials and longitudinal evaluations that systematically measure changes in clinical symptoms also remains very limited. The limitations of this evidence become increasingly relevant when the use of LLMs is confronted with risks such as crisis assessment, epistemic bias, value mismatches, and relational illusions—all of which indicate that linguistic equivalence does not necessarily reflect therapeutic legitimacy. Based on a cross-domain synthesis, this article proposes a conceptual convergence framework and a risk-stratification-based governance model to evaluate the clinical role of generative systems in psychotherapy. Currently available evidence tends to support the application of LLMs in low-risk contexts and hybrid models with professional supervision, whereas autonomous deployment in high-risk cases still requires stronger clinical evidence to obtain scientific justification.

Received: 15 Juni 2026

Revised: 30 Juli 2026

Accepted: 20 Agustus 2026

Published: 30 Agustus 2026



Copyright: © 2026 by the authors.

License Universitas Harapan Bangsa, Purwokerto, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Kata kunci: Model Bahasa Besar (LLM); Kecerdasan Buatan Generatif ; Kesehatan Mental; Aliansi Terapi Digital; Interaksi Manusia-Kecerdasan Buatan;

Pendahuluan

Perkembangan *large language models* (LLM) dan *generative AI* dalam beberapa tahun terakhir telah merekonfigurasi pola interaksi manusia-mesin, termasuk dalam layanan kesehatan mental. Model berbasis *transformer*, seperti GPT, Gemini, dan Claude, mampu menghasilkan dialog yang koheren dan kontekstual dengan karakteristik komunikasi yang menyerupai interaksi terapeutik (Stade et al., 2024). Kapabilitas mendorong pemanfaatan LLM dalam berbagai konteks psikoterapi, mulai dari psikoedukasi dan dukungan emosional mandiri hingga pendampingan dalam praktik klinis formal (Omar et al., 2024). Kecepatan adopsi teknologi tersebut melampaui perkembangan kerangka evaluasi klinis dan etis yang diperlukan untuk memastikan implementasi secara bertanggung jawab.

Integrasi *generative AI* dalam psikoterapi berkembang di tengah tekanan struktural pada layanan kesehatan mental global yang mencakup keterbatasan tenaga profesional, ketimpangan akses geografis, dan peningkatan prevalensi gangguan mental (Wang et al., 2025). LLM menawarkan skalabilitas, ketersediaan layanan secara berkelanjutan, dan efisiensi ekonomi yang berpotensi memperluas akses terhadap dukungan kesehatan mental (Blease & Torous, 2023). Keunggulan operasional belum sepenuhnya mencerminkan kesiapan klinis dan keamanan implementasi, khususnya pada konteks yang melibatkan kerentanan psikologis atau kondisi krisis (Wang et al., 2025; AlMakinah et al., 2024).

Psikoterapi merupakan praktik berisiko tinggi yang bertumpu pada relasi interpersonal, regulasi emosi dua arah, pembentukan aliansi terapeutik, serta *epistemic trust* (Nesse, 2023). Persoalan utama tidak hanya berkaitan dengan kemampuan LLM untuk mensimulasikan empati melalui bahasa, tetapi juga menyangkut kelayakan integrasinya ke dalam praktik yang secara tradisional bergantung pada kehadiran manusia dan regulasi relasional yang bersifat *embodied*. Literatur menunjukkan temuan yang beragam mengenai aspek tersebut. Sejumlah bukti melaporkan persepsi manfaat dan pembentukan aliansi terapeutik digital, sedangkan temuan lain mengidentifikasi risiko bias diagnostik, konstruksi prognostik yang bermasalah, dan keterbatasan dalam asesmen krisis (Voultsiou & Moussiades, 2026). Inovasi teknis melalui *retrieval-augmented generation*, *fine-tuning* spesifik domain, dan arsitektur multiagen telah diarahkan untuk meningkatkan kontrol serta reliabilitas respons, sementara perkembangan tersebut belum disertai validasi klinis longitudinal yang sebanding (R. Zhang et al., 2025).

Kesenjangan memerlukan pemetaan terstruktur yang tidak hanya mendeskripsikan pola pemanfaatan LLM dalam psikoterapi, tetapi juga menganalisis keterkaitan antara efektivitas klinis, keselamatan dan tata kelola, serta dinamika relasional manusia AI (Voultsiou & Moussiades, 2026; Na et al., 2025). Integrasi ketiga dimensi tersebut diperlukan untuk menghindari fragmentasi diskursus antara optimisme terhadap kapabilitas teknologi dan kehati-hatian terhadap konsekuensi klinis maupun etis. Sintesis lintas domain dapat memberikan landasan analitis untuk mengidentifikasi batas operasional, kondisi penggunaan, dan legitimasi klinis sistem generatif dalam praktik psikoterapi (R. Zhang et al., 2025).

Artikel ini menyajikan sintesis konseptual lintas domain untuk menjelaskan konvergensi LLM, *generative AI*, dan praktik psikoterapi dalam konteks klinis kontemporer. Pendekatan analitis yang digunakan mengintegrasikan bukti mengenai efektivitas klinis, keselamatan, dinamika relasional, dan desain arsitektural ke dalam suatu kerangka terpadu. Kerangka konvergensi yang diusulkan memosisikan legitimasi terapeutik LLM sebagai hasil interaksi antara kinerja klinis yang terukur, mitigasi risiko struktural, dan koherensi relasional pada konteks penggunaan tertentu. Sintesis juga menyoroti ketimpangan antara percepatan inovasi teknis dan keterbatasan validasi klinis jangka panjang serta mengusulkan model tata kelola berbasis stratifikasi risiko sebagai pendekatan

translasi yang lebih proporsional. Perspektif menggeser diskursus AI dalam psikoterapi dari persoalan kemampuan simulasi linguistik menuju evaluasi multidimensional terhadap prasyarat legitimasi klinis, batas ontologis sistem generatif, dan kebutuhan pengawasan profesional yang eksplisit.

Pendekatan Metodologis dan Sintesis Tematik

Artikel ini menggunakan pendekatan *integrative narrative review* untuk membangun sintesis konseptual lintas domain mengenai konvergensi *large language models* (LLM), *generative AI*, dan praktik psikoterapi. Pendekatan dipilih karena literatur yang tersedia menunjukkan heterogenitas metodologis yang mencakup eksperimen berbasis *vignette*, desain observasional, evaluasi arsitektur teknis, pendekatan *mixed-method*, dan analisis konseptual. Keragaman desain, tujuan, serta luaran yang dievaluasi membatasi kelayakan agregasi kuantitatif dan pelaksanaan meta-analisis yang bermakna. Kajian ini berfokus pada integrasi berbagai dimensi epistemik yang membentuk legitimasi klinis sistem generatif dalam konteks kesehatan mental, bukan pada estimasi ukuran efek suatu intervensi.

Tabel 1. Keterbatasan Konseptual dan Ruang Lingkup Penelitian

Fokus Inklusi	Batasan Eksklusi
LLM, generative AI, model berbasis transformer	Model machine learning non-generatif tanpa komponen LLM
Psikoterapi, kesehatan mental, dukungan emosional, asesmen psikologis	Domain non-kesehatan mental
Studi empiris, protokol implementasi, evaluasi arsitektural, analisis konseptual dengan implikasi klinis	Editorial, opini, atau tulisan tanpa relevansi klinis/relasional
Membahas efektivitas, keselamatan, bias, aliansi, atau model implementasi	Tidak memiliki relevansi klinis atau relasional
Publikasi 2020–2026	Di luar rentang waktu tersebut

Penelusuran literatur dilakukan menggunakan pendekatan *concept-driven* untuk mengidentifikasi publikasi yang memiliki relevansi klinis, relasional, dan arsitektural terhadap pemanfaatan LLM dalam psikoterapi. Proses pencarian mencakup tiga basis data utama, yaitu ScienceDirect (n = 579), PubMed (n = 9), dan IEEE Xplore (n = 32), dengan total 620 rekaman yang teridentifikasi pada tahap awal. Penyaringan berdasarkan judul dan penghapusan duplikasi menghasilkan 29 artikel yang dilanjutkan ke tahap penilaian abstrak. Evaluasi berikutnya mengidentifikasi 22 artikel yang memenuhi kriteria konseptual untuk penilaian teks penuh. Dari jumlah tersebut, 17 artikel dipertahankan sebagai sumber analitis utama karena memberikan kontribusi substantif terhadap pengembangan kerangka lintas domain. Artikel yang dikeluarkan pada tahap seleksi umumnya tidak memiliki relevansi klinis secara langsung, tidak berfokus pada sistem berbasis LLM, atau tidak memberikan kontribusi yang memadai terhadap pembahasan efektivitas, keselamatan, bias, dinamika relasional, maupun model *deployment*. Rentang publikasi dibatasi pada periode 2020–2026 untuk merepresentasikan perkembangan mutakhir *generative AI*. Ruang lingkup konseptual dirumuskan secara eksplisit dan disajikan dalam Tabel 1 sebagai batas analitis kajian, bukan sebagai protokol seleksi yang mengikuti kerangka *systematic review* formal.

Seleksi dilakukan secara iteratif melalui pembacaan judul, abstrak, dan teks lengkap untuk menilai kesesuaian substantif dengan kerangka analitis yang dikembangkan. Proses ini bertujuan memperoleh representasi tematik yang memadai, bukan cakupan literatur yang komprehensif secara kuantitatif. Sebanyak 17 studi dipertahankan sebagai sumber utama karena secara kolektif merepresentasikan variasi desain penelitian, konteks penggunaan, dan orientasi epistemik yang relevan dengan tujuan kajian. Karakteristik studi dirangkum dalam Tabel 2 untuk meningkatkan transparansi distribusi metodologis dan domain analisis. Heterogenitas desain mulai dari eksperimen *vignette*, studi *mixed-method*, evaluasi teknis, hingga analisis teoretis menunjukkan bahwa basis

evidensi saat ini masih bersifat eksploratif dan performatif, dengan penekanan pada persepsi kebermanfaatan dan kesetaraan struktural respons dibanding validasi klinis longitudinal. Untuk kepentingan sintesis analitis, literatur dikelompokkan ke dalam empat domain utama, meliputi efektivitas klinis dan persepsi kebermanfaatan, keselamatan dan bias epistemik, dinamika relasional manusia AI, serta arsitektur teknis dan *model deployment*.

Tabel 2. Karakteristik Studi dalam Sintesis Naratif

Author	Desain Penelitian	Konteks Penggunaan	Jenis Bukti	Domain Utama
Elyoseph & Levkovich, 2024	Eksperimen vignette komparatif	Evaluasi prognosis (skizofrenia)	Eksperimental (berbasis vignette)	Keamanan Klinis & Bias
Golden & Aboujaoude, 2024	Penerapan kerangka kerja tata kelola	Penilaian alat AI untuk OCD	Evaluatif / tata kelola	Keamanan Klinis & Tata Kelola
Hadar-Shoval et al., 2024	Studi psikometrik potong lintang	Keselaran nilai dalam kesehatan mental	Empiris (audit keselarasan)	Keamanan Klinis & Bias
Held et al., 2024	Pemodelan vignette eksperimental	Bias diagnostik (gangguan makan)	Eksperimental (analisis bias)	Keamanan Klinis & Bias
Vowels et al., 2024	Studi komparatif dengan metode campuran	Respons terapis vs. chatbot	Empiris (komparatif)	Keamanan Klinis & Risiko Relasional
Boit & Patil, 2025	Studi komputasi berskala besar	Tanya Jawab Konseling Online	Perbandingan komputasional	Efektivitas Klinis
Lozoya et al., 2025	Studi kelayakan dengan metode campuran	Sesi konseling hubungan	Empiris (interaksi yang dinilai pengguna)	Efektivitas Klinis & Dinamika Hubungan
Christiawan et al., 2026	Studi komparatif observasional	Psikoedukasi (lingkungan rawat inap)	Empiris (lingkungan klinis)	Efektivitas Klinis
Collins et al., 2025	Analisis kualitatif & analisis jaringan	Penggunaan terapi ChatGPT berbasis pendekatan naturalistik	Empiris (data naturalistik)	Dinamika Hubungan Manusia-AI
Huang et al., 2025	Survei potong lintang dengan metode campuran	Pencarian dukungan emosional	Empiris (berdasarkan survei)	Dinamika Hubungan Manusia-AI
Luo et al., 2025	Analisis konseptual teoretis	Psikoterapi Mentalisasi & Kecerdasan Buatan	Konseptual	Dinamika Hubungan Manusia-AI
Schnepper et al., 2025	Protokol kelayakan (metode campuran)	Peningkatan CBT (dialog Sokrates)	Kelayakan / implementasi hibrida	Model Penerapan Hibrida
Schulich et al., 2025	Makalah kerangka konseptual	Desain chatbot LLM untuk kesehatan mental	Teknis-konseptual	Arsitektur Teknis & Tata Kelola
Surana et al., 2025	Evaluasi eksperimental teknis	Chatbot kesehatan mental berbasis RAG	Perbandingan teknis	Arsitektur Teknis
Yirmiya & Fonagy, 2025	Evaluasi teknis multi-model	Chatbot pendukung penanggulangan masalah	Perbandingan teknis	Arsitektur Teknis
Elboardy et al., 2025	Studi pengembangan penyempurnaan	Chatbot kesehatan mental dwibahasa	Perkembangan teknis	Arsitektur Teknis
Attanasio et al., 2026	Studi pengembangan & kegunaan	Simulasi pelatihan psikoterapi	Empiris (kegunaan)	Pelatihan & Pendidikan

Kapabilitas Linguistik dan Batas Klinis LLM dalam Psikoterapi

LLM berbasis arsitektur transformer merupakan sistem generatif yang dilatih untuk memprediksi token berikutnya berdasarkan konteks linguistik pada skala data besar (OpenAI et al., 2024). Mekanisme *self-attention* memungkinkan model mempertahankan koherensi percakapan sekaligus menghasilkan respons yang adaptif terhadap variasi masukan pengguna (Lin et al., 2022). Sebagai *foundation models*, LLM memiliki karakteristik *general-purpose* yang memungkinkan adaptasi melalui *prompting*, *fine-tuning*, maupun integrasi dengan sistem pendukung. Dalam konteks psikoterapi, kemampuan tercermin melalui dialog generatif yang dapat mereplikasi

sejumlah teknik komunikasi terapeutik, seperti refleksi, validasi emosional, dan pertanyaan *Socratic* (R. Zhang et al., 2025; Guo et al., 2024).

Karakteristik *general-purpose* sekaligus membatasi fungsi LLM dalam konteks klinis (Thirunavukarasu et al., 2023). LLM pada dasarnya tidak dikembangkan secara spesifik untuk melakukan diagnosis, menjalankan regulasi emosi secara timbal balik, atau memberikan intervensi psikoterapeutik terstruktur. Kompetensi terapeutik yang ditampilkan lebih merepresentasikan kemampuan model dalam mereproduksi pola linguistik yang dipelajari daripada pemahaman klinis intrinsik. Literatur konseptual mengenai mentalisasi dan *epistemic trust* juga menunjukkan bahwa simulasi empati secara linguistik tidak dapat disetarakan dengan relasi terapeutik yang bersifat *embodied* dan melibatkan regulasi afektif timbal balik (Yirmiya & Fonagy, 2025). Kapasitas dialogis LLM lebih tepat dipahami sebagai representasi respons terapeutik yang disimulasikan, bukan sebagai agensi klinis yang otonom (Q. Zhang et al., 2025).

Keterbatasan mendorong pengembangan berbagai mekanisme pembatas untuk meningkatkan kesesuaian klinis sekaligus mengurangi risiko respons yang tidak terkendali. Pendekatan *retrieval-augmented generation* meningkatkan keterikatan respons pada sumber informasi yang terkurasi sehingga dapat memperkuat *groundedness* system (Boit & Patil, 2025). Arsitektur multiagen yang dilengkapi mekanisme supervisi internal dan *human-in-the-loop* juga dikembangkan untuk meningkatkan pengawasan serta membatasi keluaran yang berpotensi menimbulkan risiko (Cho et al., 2025; Li et al., 2023). Perkembangan arsitektural tersebut dapat meningkatkan stabilitas dan kontrol teknis, tetapi legitimasi klinis tetap memerlukan pembatasan domain, mekanisme deteksi risiko, dan pengawasan profesional yang eksplisit. Kompleksitas arsitektur dengan sendirinya belum dapat menjadi dasar untuk menetapkan kelayakan klinis suatu sistem generative (Blease & Rodman, 2024).

Efektivitas Klinis LLM dalam Psikoterapi

Efektivitas Persepsional, Empati, dan Keterbatasan *Clinical Outcomes*

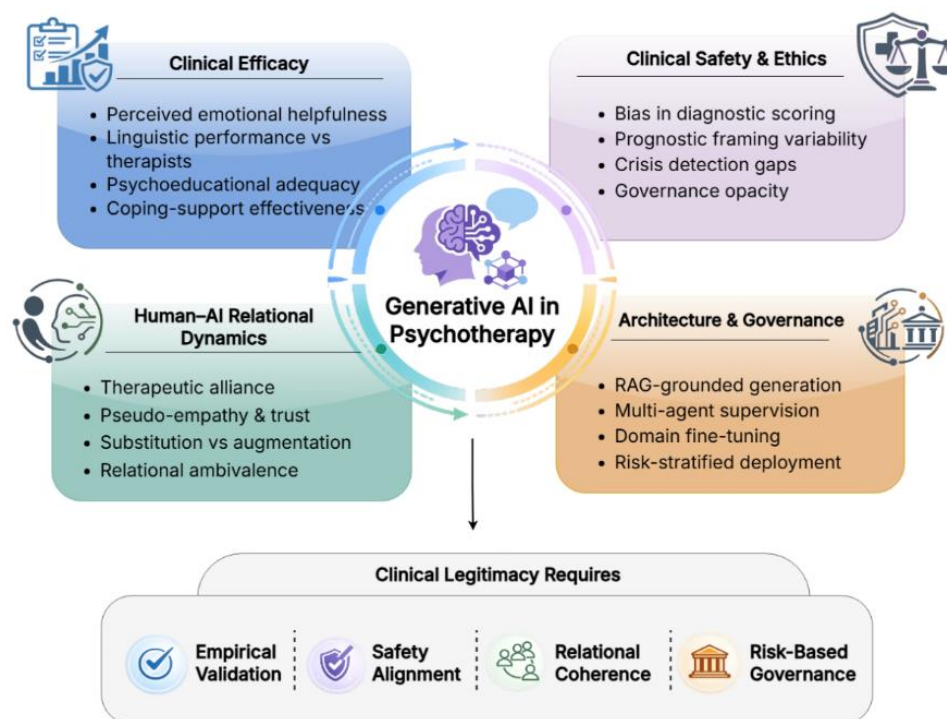
Evidensi mengenai efektivitas LLM dalam konteks psikoterapi menunjukkan bahwa sistem generatif mampu menghasilkan respons yang secara linguistik terstruktur, reflektif, dan konsisten dengan format komunikasi terapeutik manusia. Dalam konteks tanya jawab konseling daring, respons ChatGPT dinilai sebanding secara struktural dengan konselor manusia dalam organisasi argumen dan koherensi refleksi (Huang et al., 2025). Interaksi berbasis GPT-4 dalam sesi relasional tunggal memperlihatkan tingkat empati linguistik dan validasi emosional yang tinggi menurut penilaian partisipan (Vowels et al., 2024), sementara penggunaan GPT-4o sebagai pelengkap psikoedukasi pada pasien rawat inap dipersepsikan membantu profesional kesehatan mental menilai kedalaman klinisnya lebih terbatas (Attanasio et al., 2026).

Pengalaman merasa “didengar” dan “dipahami” secara konsisten menjadi determinan penting dalam membentuk persepsi kebermanfaatan sistem (Collins et al., 2025; Luo et al., 2025). Dalam konteks tersebut, empati linguistik berperan sebagai komponen utama yang mendukung evaluasi positif terhadap interaksi berbasis LLM. Kesetaraan pada tingkat bahasa dan persepsi empatik terutama mencerminkan kompetensi responsif dalam kondisi terstruktur serta kemampuan memberikan dukungan emosional jangka pendek, bukan bukti perubahan gejala klinis yang telah tervalidasi secara longitudinal. Variasi performa antar model dan perbedaan *framing* respons juga mengindikasikan bahwa konsekuensi klinis tidak melekat secara inheren pada arsitektur LLM, tetapi dipengaruhi oleh konfigurasi sistem dan konteks implementasinya (Elyoseph & Levkovich, 2024; Surana et al.,

2025). Basis bukti pada domain efektivitas saat ini lebih banyak merepresentasikan performa komunikatif dan penerimaan pengguna daripada efektivitas psikoterapeutik dalam pengertian klinis formal.

Keterbatasan Metodologis dan Kesenjangan Evidensi Klinis

Basis evidensi yang tersedia masih menghadapi keterbatasan metodologis yang substansial. Evaluasi LLM sebagai agen terapeutik belum secara luas menerapkan pengukuran *pre-post* menggunakan instrumen klinis terstandar, pemantauan longitudinal terhadap perubahan gejala, maupun *randomized controlled trials*. Sebagian besar bukti masih bertumpu pada desain observasional, *experimental vignette*, dan evaluasi sesi tunggal yang berfokus pada kualitas respons serta persepsi pengguna. Karakteristik metodologis menyebabkan klaim efektivitas lebih banyak merepresentasikan dukungan emosional jangka pendek daripada perubahan terapeutik yang tervalidasi melalui evaluasi klinis terkontrol.



Gambar 1. Pemetaan Bukti Terstruktur tentang Kecerdasan Buatan Generatif dalam Psikoterapi

Variabilitas *framing* prognostik dan bias respons antarmodel semakin menunjukkan bahwa dampak klinis bersifat kontekstual serta sensitif terhadap konfigurasi arsitektur dan kondisi implementasi (Elyoseph & Levkovich, 2024; Surana et al., 2025). Ketiadaan desain eksperimental yang ketat, pelaporan *outcomes* terstandar, dan data keselamatan prospektif membatasi interpretasi kesetaraan performatif pada tingkat linguistik sebagai indikator efektivitas psikoterapeutik dalam hierarki evidensi klinis yang mapan. Distribusi dan ketidakseimbangan basis evidensi pada berbagai domain dirangkum secara terstruktur dalam Gambar 1. Pemetaan menunjukkan dominasi evaluasi performatif dan perseptual dibandingkan validasi klinis longitudinal sekaligus memperlihatkan kesenjangan antara kemajuan teknis dan kematangan evidensi terapeutik. Kerangka visual memberikan landasan analitis untuk memosisikan bukti LLM yang tersedia saat ini dalam spektrum efektivitas klinis.

Keamanan Klinis, Risiko Struktural, dan Tata Kelola LLM dalam Psikoterapi

Risiko Klinis dan Dinamika Relasional

Kompetensi performatif LLM pada konteks tertentu belum menghilangkan risiko klinis yang muncul dalam penerapannya pada psikoterapi. Risiko terutama berkaitan dengan kemungkinan munculnya respons yang tidak akurat, tidak sesuai konteks, atau tidak memadai ketika sistem menghadapi situasi berisiko tinggi. *Chatbot* berbasis LLM mampu memberikan validasi emosional awal, tetapi masih menunjukkan keterbatasan dalam melakukan asesmen krisis dan merespons risiko bunuh diri secara tepat (Scholich et al., 2025). Evaluasi terhadap GPT-4 juga menunjukkan kemampuan menghasilkan respons empatik dalam intervensi relasional, sementara eksplorasi risiko dan perumusan rencana keselamatan masih memiliki keterbatasan (Vowels et al., 2024).

Risiko penggunaan LLM juga mencakup dimensi relasional yang dapat memengaruhi pola interaksi pengguna dengan sistem. Sejumlah pengguna menjadikan LLM sebagai sumber dukungan utama ketika akses terhadap layanan terapi profesional terbatas (Luo et al., 2025). Persepsi bahwa sistem dapat “berfungsi seperti terapis” berpotensi meningkatkan keterlibatan pengguna sekaligus membentuk ketergantungan emosional dan ilusi relasional (Collins et al., 2025). Konsep *pseudo-empathy* menunjukkan bahwa simulasi empati tanpa regulasi afektif timbal balik dapat membentuk *epistemic trust* yang tidak sepenuhnya berlandaskan realitas relasional klinis, terutama pada individu dengan tingkat kerentanan yang tinggi (Yirmiya & Fonagy, 2025). Risiko LLM dalam psikoterapi tidak terbatas pada kesalahan faktual atau *hallucination*, tetapi juga mencakup dinamika relasional yang dapat memengaruhi ekspektasi terapeutik, regulasi emosi, dan persepsi keamanan psikologis pengguna.

Bias Epistemik dan Ketidakselarasan Nilai

Literatur mengidentifikasi bias epistemik dan ketidakselarasan nilai sebagai faktor yang berpotensi memengaruhi proses terapeutik berbasis LLM. Bias gender teridentifikasi ketika LLM mengevaluasi *vignette* gangguan makan, dengan partisipan laki-laki memperoleh penilaian kualitas hidup mental yang lebih buruk tanpa dukungan empiris yang memadai (Schnepper et al., 2025). Temuan menunjukkan potensi sistem untuk mereproduksi stereotip sosial dalam proses asesmen psikologis. Variasi *framing* prognostik antarmodel juga ditemukan dalam konteks skizofrenia, dengan beberapa model menghasilkan proyeksi yang lebih pesimistis dibandingkan model lainnya (Elyoseph & Levkovich, 2024). Variasi berpotensi memengaruhi pembentukan harapan terapeutik dan motivasi pengguna selama interaksi klinis.

LLM juga menunjukkan orientasi nilai yang relatif konsisten terhadap *universalism* dan *self-direction*, yang berpotensi menghasilkan ketidakselarasan dengan nilai sosial dan budaya pada kelompok pengguna tertentu (Hadar-Shoval et al., 2024). Dalam konteks psikoterapi, ketidakselarasan dapat memengaruhi pembentukan aliansi terapeutik sekaligus meningkatkan risiko *epistemic influence*, yaitu kondisi ketika respons sistem secara implisit membentuk preferensi, interpretasi, atau kerangka pemaknaan pengguna. Persoalan bias pada LLM dengan demikian melampaui dimensi keadilan teknis karena turut mencakup aspek epistemologis yang berkaitan dengan pembentukan makna, harapan, dan interpretasi klinis.

Tata Kelola dan Mitigasi Risiko

Risiko klinis dan epistemik dalam penggunaan LLM memiliki keterkaitan erat dengan tata kelola sistem generatif. Evaluasi menggunakan kerangka FAITA-Mental Health menunjukkan bahwa platform *generative AI* dalam

kesehatan mental masih menghadapi keterbatasan terkait transparansi data, kejelasan kebijakan privasi, dan dokumentasi mekanisme respons terhadap kondisi krisis (Golden & Aboujaoude, 2024). Keterbatasan berimplikasi terhadap *informed consent*, keamanan informasi, dan akuntabilitas profesional dalam implementasi klinis. Adaptasi teknis melalui *fine-tuning* spesifik domain juga belum secara otomatis mengatasi bias kurasi, keterbatasan representasi data, maupun kebutuhan validasi klinis (Elboardy et al., 2025). Ketidakjelasan standar regulasi mengenai tanggung jawab dan *liability* dalam interaksi AI-pasien semakin memperluas ambiguitas institusional dalam penerapan sistem berbasis LLM.

Berbagai mekanisme mitigasi arsitektural telah dikembangkan untuk meningkatkan kontrol, *groundedness*, dan keamanan keluaran sistem. Pendekatan *retrieval-augmented generation*, arsitektur multiagen dengan mekanisme *human-in-the-loop*, serta *guardrails* berbasis *prompt engineering* dapat berfungsi sebagai lapisan protektif untuk mengurangi *hallucination* dan membatasi keluaran yang berpotensi menimbulkan risiko (Held et al., 2024; Christiawan et al., 2026; Surana et al., 2025). Kerangka *prompt engineering* terstruktur dengan *guardrails* eksplisit juga dapat berfungsi sebagai *containment layer* untuk membatasi variasi respons yang tidak sesuai dengan konteks klinis (Boit & Patil, 2025). Mekanisme teknis tetap memerlukan supervisi klinis dan kerangka regulasi formal untuk menjamin penerapan yang bertanggung jawab. Tata kelola LLM dalam psikoterapi memerlukan pendekatan berlapis yang mengintegrasikan audit bias secara sistematis, transparansi sistem, mekanisme eskalasi krisis, perlindungan data, serta delimitasi tanggung jawab profesional. Kerangka tata kelola yang disertai stratifikasi risiko diperlukan untuk menyesuaikan tingkat pengawasan dengan karakteristik penggunaan dan potensi konsekuensi klinis. Ketiadaan mekanisme dapat meningkatkan paparan terhadap kesalahan klinis, risiko *liability*, dan penurunan kepercayaan terhadap institusi yang mengintegrasikan sistem generatif ke dalam layanan kesehatan mental.

Dinamika Relasional dan Epistemik dalam Psikoterapi Manusia-AI

Aliansi Terapeutik dan *Epistemic Trust* dalam Interaksi Manusia

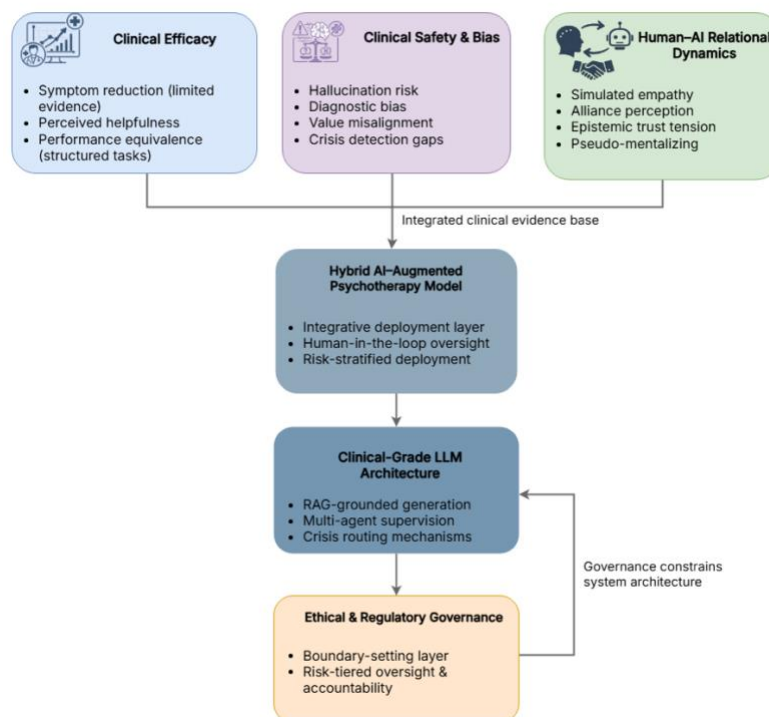
Aliansi terapeutik dalam interaksi berbasis LLM tidak terbentuk melalui relasi interpersonal yang bersifat *embodied*, tetapi melalui konfigurasi sinyal linguistik yang diasosiasikan dengan empati, refleksi, dan responsivitas. Validasi emosional, parafrase reflektif, dan konsistensi dialogis berfungsi sebagai penanda relasional yang membentuk persepsi pengguna terhadap kualitas interaksi. Pengalaman merasa “didengar” dan “dipahami” menjadi determinan penting dalam evaluasi positif terhadap ChatGPT (Luo et al., 2025), sedangkan persepsi bahwa sistem dapat “berfungsi seperti terapis” menjadi salah satu elemen penting dalam dinamika keterlibatan pengguna (Collins et al., 2025). Aliansi yang terbentuk dalam konteks tersebut bersifat performatif dan perseptual karena muncul melalui koherensi linguistik serta konsistensi respons, bukan melalui regulasi emosi timbal balik atau keterlibatan afektif dua arah.

Kemampuan berkaitan dengan kapasitas LLM untuk mereplikasi pola bahasa reflektif yang menyerupai proses mentalisasi. Sistem generatif tidak memiliki pengalaman subjektif maupun kapasitas mentalisasi yang bersifat *embodied* sebagaimana terdapat dalam relasi interpersonal manusia (Yirmiya & Fonagy, 2025). Perbedaan ontologis membedakan simulasi empati dari relasi terapeutik yang melibatkan regulasi afektif bersama dan pembentukan *epistemic trust* melalui pengalaman interpersonal. Dalam praktiknya, pengguna dapat mengatribusikan agensi dan intensionalitas kepada sistem AI, membentuk ekspektasi relasional tertentu, bahkan menggunakan sistem sebagai substitusi sementara bagi terapi manusia ketika akses terhadap layanan profesional terbatas (Collins et al., 2025; Luo et al., 2025). Pola interaksi dapat meningkatkan keterlibatan pengguna, tetapi

juga berpotensi mendorong *over-reliance* dan ilusi relasional, terutama pada kondisi *distress* berat. Variasi *framing* dan gaya respons antarmodel semakin menunjukkan bahwa pembentukan aliansi dan kepercayaan tidak hanya dipengaruhi oleh karakteristik pengguna, tetapi juga oleh konfigurasi sistem. Interaksi manusia AI dengan demikian memiliki dimensi *epistemic influence* karena respons yang dihasilkan berpotensi memengaruhi cara pengguna membentuk interpretasi, ekspektasi, dan pemaknaan terhadap pengalaman psikologisnya (Elyoseph & Levkovich, 2024).

Legitimasi Klinis di Luar Kesenjangan Linguistik

Perbandingan langsung antara respons LLM dan terapis manusia menunjukkan kesetaraan struktural pada level organisasi bahasa, kejelasan refleksi, dan konsistensi format respons (Huang et al., 2025). Studi komparatif juga mengidentifikasi perbedaan konsisten dalam kedalaman eksplorasi klinis, fleksibilitas interpretatif, dan sensitivitas terhadap dinamika risiko (Scholich et al., 2025; Vowels et al., 2024). Kesenjangan linguistik tidak identik dengan kesetaraan klinis, khususnya dalam pengambilan keputusan kompleks atau situasi distress tinggi. Legitimasi terapeutik LLM perlu dipahami sebagai konstruk multidimensional yang mencakup kompetensi responsif, efektivitas yang terukur, mitigasi risiko, dan koherensi relasional dalam konteks penggunaan tertentu. Kerangka konvergensi pada Gambar 2 berfungsi sebagai *evaluative lens* untuk memetakan posisi klinis sistem AI berdasarkan keterkaitan antardimensi tersebut. Kerangka ini mengintegrasikan bukti mengenai efektivitas, kontrol terhadap risiko struktural, dan kualitas relasional yang dipersepsikan pengguna sebagai dasar evaluasi terhadap kelayakan penggunaan LLM dalam konteks klinis. Perspektif menggeser fokus evaluasi dari keserupaan linguistik dengan terapis manusia menuju penilaian multidimensional terhadap prasyarat legitimasi klinis berdasarkan karakteristik dan tingkat risiko penggunaan.



Gambar 2. Kerangka Kerja Konvergensi Evaluatif untuk Menilai Keabsahan Klinis Sistem Psikoterapi Berbasis LLM

Distingsi Konseptual dalam Psikoterapi Berbasis AI

Diskursus mengenai LLM dalam psikoterapi masih menghadapi persoalan konseptual dalam membedakan simulasi linguistik dari agensi terapeutik. Kemampuan sistem untuk menghasilkan respons reflektif dapat dipersepsikan sebagai representasi kompetensi klinis, meskipun mekanisme generatifnya bertumpu pada pemodelan statistik pola bahasa dan tidak melibatkan mentalisasi timbal balik maupun tanggung jawab profesional sebagaimana terdapat dalam praktik psikoterapi manusia (Yirmiya & Fonagy, 2025; Thirunavukarasu et al., 2023; Blease & Rodman, 2024). Peningkatan performa komunikatif tidak identik dengan kompetensi klinis tervalidasi (Wang et al., 2025; Guo et al., 2024). Distingsi serupa diperlukan antara persepsi empati dan efektivitas terapeutik, serta antara skalabilitas teknis dan kesiapan klinis. Pengguna dapat mempersepsikan interaksi dengan LLM sebagai empatik dan membantu (Vowels et al., 2024; Luo et al., 2025), sedangkan keterbatasan validasi longitudinal, pelaporan keselamatan prospektif, dan standardisasi *clinical outcomes* masih membatasi interpretasi terhadap legitimasi terapeutiknya (Wang et al., 2025; Q. Zhang et al., 2025). Evaluasi komparatif juga mengidentifikasi keterbatasan dalam asesmen risiko dan kedalaman eksplorasi klinis (Scholich et al., 2025). Dukungan emosional yang dipersepsikan pengguna dan kemampuan menjangkau populasi secara luas belum dapat diperlakukan sebagai indikator yang setara dengan legitimasi terapeutik. Validasi klinis terstandar dan tata kelola risiko yang proporsional tetap diperlukan agar perluasan penggunaan LLM tidak melampaui tingkat kematangan evidensi yang tersedia.

Model Penerapan LLM dalam Psikoterapi dan Pendidikan Klinis

Spektrum Penerapan: Model Otonom dan Hibrida

Penerapan LLM dalam psikoterapi berada pada spektrum yang mencakup konfigurasi otonom hingga model hibrida dengan supervisi profesional. Pada konfigurasi otonom, sistem berfungsi sebagai agen interaktif tanpa keterlibatan langsung klinisi dan menawarkan skalabilitas serta aksesibilitas pada kondisi dengan keterbatasan layanan profesional. Sejumlah pengguna memanfaatkan ChatGPT sebagai sumber dukungan emosional mandiri atau substitusi sementara ketika akses terhadap terapi profesional terbatas (Luo et al., 2025; Collins et al., 2025). Dalam konteks konseling daring yang terstruktur, respons yang dihasilkan juga menunjukkan karakteristik linguistik yang sebanding dengan respons konselor manusia (Huang et al., 2025). Keterbatasan nyata muncul pada asesmen risiko, respons krisis, dan kedalaman eksplorasi klinis (Vowels et al., 2024)(Scholich et al., 2025). Variabilitas *framing* antarmodel turut menimbulkan risiko epistemik karena respons yang berbeda berpotensi memengaruhi ekspektasi dan interpretasi pengguna terhadap kondisi psikologisnya (Elyoseph & Levkovich, 2024).

Konfigurasi otonom memiliki dasar evidensi yang lebih kuat untuk fungsi berisiko rendah, seperti psikoedukasi dan dukungan emosional ringan, dibandingkan untuk psikoterapi intensif yang memerlukan asesmen klinis, pengambilan keputusan kompleks, serta manajemen krisis. Penerapan pada konteks dengan risiko lebih tinggi memerlukan mekanisme pembatasan, deteksi risiko, eskalasi, dan pengawasan profesional yang memadai. Pada spektrum yang lebih terkendali, model hibrida memosisikan LLM sebagai instrumen augmentatif dalam kerangka supervisi profesional. Sistem AI dapat berfungsi sebagai fasilitator refleksi, pendukung psikoedukasi, atau instrumen untuk menjalankan tugas terstruktur, sedangkan tanggung jawab terhadap keputusan klinis tetap berada pada tenaga profesional. Implementasi Socrates 2.0 dalam konteks *cognitive behavioral therapy* (CBT) menunjukkan pemanfaatan desain multiagen dan supervisi internal untuk mendukung dialog *Socratic* sebagai aktivitas di luar sesi tanpa menggantikan intervensi terapeutik utama (Held et al., 2024). Pemanfaatan GPT-4o

sebagai pelengkap psikoedukasi pada pasien rawat inap juga menunjukkan potensi peningkatan persepsi empati tanpa mengambil alih fungsi klinis inti (Attanasio et al., 2026). Evaluasi tata kelola menekankan pentingnya transparansi peran AI, delimitasi tanggung jawab profesional, dan integrasi dalam protokol klinis formal (Golden & Aboujaoude, 2024). Berdasarkan tingkat kematangan evidensi yang tersedia, model hibrida memberikan mekanisme pengawasan yang lebih selaras dengan keterbatasan fungsional LLM dan kebutuhan keselamatan klinis dibandingkan konfigurasi yang menggantikan keterlibatan profesional secara penuh.

LLM dalam Pelatihan dan Pendidikan Klinis

Pemanfaatan LLM dalam pelatihan dan pendidikan profesional kesehatan mental merepresentasikan use case dengan profil risiko paling rendah serta potensi manfaat yang substansial. GPT-4 dapat digunakan sebagai klien simulasi dalam pelatihan psikoterapi dengan tingkat realisme moderat dan tingkat penerimaan yang positif di kalangan profesional (Lozoya et al., 2025). Dalam konteks tersebut, LLM memungkinkan peserta pelatihan menghadapi beragam skenario klinis tanpa memberikan intervensi secara langsung kepada pasien serta menyediakan ruang untuk refleksi dan umpan balik yang terstruktur. Lingkungan pendidikan juga memungkinkan eksplorasi sistematis terhadap pola komunikasi terapeutik, respons klinis, dan pembentukan aliansi dalam kondisi yang lebih terkendali. Pelatihan profesional sebagai konteks yang relevan untuk mengevaluasi kapabilitas dan keterbatasan model sebelum penerapan yang lebih luas pada populasi klinis. Berdasarkan bukti yang tersedia, pemanfaatan LLM untuk simulasi dan pengembangan kompetensi profesional menunjukkan kesiapan implementasi yang lebih besar dibandingkan penggunaan otonom sebagai agen terapeutik. Dalam kerangka translasi klinis, domain pendidikan dapat berfungsi sebagai lingkungan evaluatif untuk menguji kapasitas dialogis, konsistensi respons, dan batas operasional LLM tanpa memberikan paparan risiko terapeutik secara langsung kepada pasien.

Kematangan Evidensi dan Kesiapan Translasi Klinis

Distribusi evidensi lintas domain pada Tabel 3 menunjukkan tingkat kematangan yang tidak seragam. Domain arsitektur teknis memperlihatkan perkembangan dalam pengendalian respons dan mitigasi *hallucination* melalui pendekatan seperti *retrieval-augmented generation* dan desain multiagen (Christiawan et al., 2026; Held et al., 2024; Surana et al., 2025). Sebaliknya, bukti mengenai efektivitas klinis masih didominasi oleh desain observasional dan evaluasi sesi tunggal dengan dukungan validasi longitudinal yang terbatas (Vowels et al., 2024; Attanasio et al., 2026). Pada domain relasional dan bias, temuan menunjukkan kompleksitas dinamika epistemik yang belum sepenuhnya terintegrasi dalam desain arsitektural (Schnepper et al., 2025; Luo et al., 2025). Ketimpangan menunjukkan bahwa integrasi LLM dalam psikoterapi tidak hanya bergantung pada peningkatan performa teknis, tetapi juga memerlukan proses translasi klinis yang mengintegrasikan efektivitas terukur, mitigasi risiko, dan koherensi relasional. Tingkat kematangan pada setiap dimensi menjadi dasar untuk membedakan konteks penggunaan yang sesuai dengan fungsi augmentatif, aplikasi berisiko rendah, dan penggunaan otonom pada populasi klinis berisiko tinggi. Pemetaan komparatif pada Tabel 3 merangkum ketimpangan dan memberikan dasar analitis untuk menilai kesiapan translasi LLM berdasarkan karakteristik serta tingkat risiko penggunaannya.

Tabel 3. Pemetaan Bukti Komparatif Lintas Bidang

Domain	Desain-Desain Unggulan	Karakterisasi Bukti	Keterbatasan Utama	Studies
Efektivitas Klinis	Observasional; metode campuran; komparatif	Persepsi tentang kegunaan; kesetaraan respons struktural (eksploratif)	Tidak ada uji klinis terkontrol acak (RCT); tidak ada pemantauan gejala secara longitudinal; sampel kecil	(Luo et al., 2025; Vowels et al., 2024; Attanasio et al., 2026; Huang et al., 2025)
Keamanan Klinis & Bias	Eksperimen vignette; audit keselarasan	Bias diagnostik; risiko halusinasi; ketidaksesuaian nilai (berbasis model)	Tidak ada uji klinis dalam situasi krisis di dunia nyata; tidak ada data prospektif mengenai risiko	(Schnepper et al., 2025; Elyoseph & Levkovich, 2024; Golden & Aboujaoude, 2024 ; Hadar-Shoval et al., 2024
Dinamika Hubungan Manusia-AI	Metode campuran; kualitatif; teoretis	Persepsi aliansi; simulasi empati; kepercayaan epistemik (konseptual-empiris)	Aliansi yang diukur secara perseptual; tanpa validasi yang didasarkan pada pengalaman tubuh	(Yirmiya & Fonagy, 2025;Collins et al., 2025; Scholich et al., 2025; Vowels et al., 2024)
Arsitektur Teknis & Penerapan	Perbandingan kinerja; desain multi-agen; penyempurnaan	Kestabilan; pengurangan halusinasi; pengendalian sistem (teknis)	Validasi hasil klinis yang terbatas; adanya kesenjangan dalam pengujian ekologi	(Christiawan et al., 2026 ;Surana et al., 2025; Elboardy et al., 2025 ; Held et al., 2024)

Arah Pengembangan dan Translasi Klinis LLM

Pengembangan Arsitektur LLM Berorientasi Klinis

Literatur menunjukkan bahwa sebagian besar LLM yang digunakan dalam konteks psikoterapi masih merupakan model general-purpose yang diadaptasi, bukan arsitektur yang sejak awal dirancang untuk penggunaan klinis (Jesudason et al., 2025). Pendekatan *retrieval-augmented generation* dan desain multiagen dengan mekanisme supervisi internal menunjukkan potensi dalam meningkatkan *groundedness* serta membatasi keluaran yang berisiko (Held et al., 2024; Christiawan et al., 2026). Penerapan *prompt engineering* secara terstruktur juga menunjukkan bahwa pembatasan instruksional dan *system guardrails* dapat memperkuat *boundary-setting* dalam dialog yang menyerupai interaksi terapeutik (Boit & Patil, 2025). Peningkatan kontrol respons melalui desain arsitektural belum dapat disetarakan dengan kompetensi klinis atau validasi terapeutik jangka panjang (Fanarioti & Karpouzis, 2025). Kemajuan teknis juga memerlukan evaluasi lebih lanjut untuk menentukan keterkaitannya dengan keselamatan klinis dan kualitas relasional. Pengembangan arsitektur *clinical-grade* perlu mengadopsi prinsip *domain-constrained* dan *supervision-aware* melalui pembatasan konteks penggunaan, mekanisme deteksi kondisi krisis, serta integrasi *human-in-the-loop* secara eksplisit (Grodniewicz & Hohol, 2023). Orientasi pengembangan perlu melampaui akurasi linguistik dengan mempertimbangkan stabilitas respons pada kondisi *distress* tinggi, sensitivitas terhadap konteks klinis, dan transparansi operasional. Integrasi kebutuhan klinis ke dalam desain sistem menjadi prasyarat penting agar peningkatan performa generatif dapat berkembang menuju kapabilitas yang relevan secara terapeutik.

Standardisasi Evaluasi dan Validasi Klinis

Literatur yang tersedia masih menghadapi keterbatasan berupa belum konsistennya standar evaluasi klinis untuk menilai sistem berbasis LLM. Sebagian besar evaluasi mengandalkan *self-report*, persepsi empati, atau analisis linguistik, sedangkan bukti longitudinal dengan pengukuran gejala menggunakan instrumen terstandar masih terbatas (Luo et al., 2025 ; Vowels et al., 2024). Perbedaan evaluasi antara pasien dan tenaga profesional semakin menegaskan kebutuhan terhadap instrumen yang mampu mengintegrasikan *clinical outcomes*, kualitas relasional, dan indikator keselamatan secara komprehensif (Attanasio et al., 2026). Keterbatasan desain

eksperimental dan heterogenitas pelaporan luaran menyebabkan basis evidensi yang tersedia masih didominasi oleh temuan eksploratif dibandingkan bukti intervensional yang telah tervalidasi. Kerangka evaluasi selanjutnya perlu mencakup uji klinis terkontrol, pengukuran *pre-post* menggunakan instrumen klinis terstandar untuk menilai perubahan gejala, pemantauan longitudinal, audit bias secara sistematis (Schnepper et al., 2025), serta penilaian keselarasan nilai budaya (Hadar-Shoval et al., 2024). Pelaporan insiden keselamatan dan mekanisme eskalasi klinis juga perlu diintegrasikan sebagai komponen evaluasi, bukan hanya sebagai atribut teknis sistem. Standardisasi metrik dan transparansi pelaporan diperlukan untuk meningkatkan komparabilitas antarevaluasi sekaligus memperkuat dasar evidensi bagi penilaian efektivitas dan keselamatan klinis.

Tata Kelola Berbasis Risiko dan Integrasi Klinis

Mengingat keterbatasan LLM yang bersifat tidak hanya teknis tetapi juga ontologis khususnya ketiadaan mentalisasi timbal balik dan regulasi emosi embodied legitimasi klinis tidak dapat diperlakukan sebagai atribut inheren sistem generatif (Yirmiya & Fonagy, 2025). Evaluasi tata kelola menunjukkan bahwa transparansi data, mekanisme eskalasi krisis, serta delimitasi akuntabilitas profesional masih belum terstandarisasi secara memadai(Golden & Aboujaoude, 2024). Variabilitas *framing* prognostik, bias diagnostik antarmodel, dan ketidakselarasan nilai semakin menunjukkan bahwa risiko klinis maupun epistemik bersifat kontekstual serta dipengaruhi oleh konfigurasi sistem dan kondisi implementasinya (Schnepper et al., 2025; Elyoseph & Levkovich, 2024; Hadar-Shoval et al., 2024). Tata kelola berbasis stratifikasi risiko menawarkan kerangka translasi yang lebih proporsional untuk mengakomodasi variasi tersebut. Pendekatan ini menyesuaikan tingkat supervisi, persyaratan validasi klinis, dan kewajiban akuntabilitas dengan intensitas keterlibatan sistem dalam proses terapeutik serta potensi dampaknya terhadap pengguna. Kerangka yang dirangkum dalam Tabel 4 memosisikan integrasi LLM sebagai proses bertahap yang bergantung pada konvergensi antara kesiapan arsitektur teknis, kekuatan validasi empiris, dan struktur pengawasan profesional. Stratifikasi memungkinkan aplikasi berisiko rendah memperoleh tingkat pengawasan yang proporsional, sedangkan penggunaan pada konteks berisiko tinggi memerlukan validasi dan supervisi yang lebih ketat. Ketiadaan pendekatan berbasis risiko dapat menghasilkan ketidakseimbangan regulasi, baik melalui pembatasan berlebihan terhadap aplikasi berisiko rendah maupun pengawasan yang tidak memadai terhadap penggunaan berisiko tinggi. Ketidakseimbangan berpotensi mengaburkan tanggung jawab klinis dan melemahkan legitimasi institusional dalam penerapan LLM pada layanan psikoterapi.

Tabel 4. Kerangka Stratifikasi Risiko untuk Tata Kelola LLM dalam Psikoterapi

Kasus Penggunaan	Tingkat Risiko Klinis	Dasar Bukti yang Umum	Pengawasan yang Direkomendasikan	Persyaratan Tata Kelola
Chatbot Psikoedukasi / Informasi	Rendah	Studi observasional & studi kegunaan	Peninjauan berkala oleh manusia	Transparansi & perlindungan data
Dukungan Penanganan / Asisten Bantuan Mandiri	Rendah-Sedang	Laporan pengguna & metode campuran	Penanganan krisis + jalur eskalasi	Audit bias + batasan tanggung jawab yang jelas
Peningkatan CBT (model hibrida)	Sedang	Studi kelayakan + studi arsitektur	Pengawasan dengan keterlibatan manusia	Integrasi protokol klinis + dokumentasi
Penilaian diagnostik / Kerangka prognostik	Sedang-Tinggi	Studi eksperimental vignette	Diperlukan pengawasan oleh pihak yang ahli	Validasi bias + audit model
Agen terapeutik yang sepenuhnya otonom	Tinggi	Tidak ada uji klinis terkontrol acak (RCT); hanya bukti eksploratif	Pengawasan klinis berkelanjutan	Tinjauan regulasi independen + pelaporan keselamatan

Kesimpulan

Sebagai *narrative synthesis*, kajian ini tidak dirancang sebagai tinjauan sistematis yang mencakup keseluruhan literatur yang tersedia. Pemilihan publikasi didasarkan pada relevansi konseptual dan klinis terhadap kerangka analitis yang dikembangkan, bukan pada protokol *systematic review* formal. Konsekuensinya, evaluasi *risk of bias* secara terstruktur dan penilaian kualitas metodologis pada setiap publikasi tidak diterapkan. Literatur yang disintesis juga menunjukkan heterogenitas metodologis yang substansial, mencakup eksperimen berbasis *vignette*, survei, evaluasi teknis, dan analisis konseptual. Keragaman desain, tujuan, serta luaran tersebut membatasi komparabilitas langsung antarpublikasi dan tidak memungkinkan agregasi kuantitatif yang bermakna.

Penelusuran literatur dibatasi pada tiga basis data utama dan tidak mencakup *grey literature* maupun laporan industri yang belum melalui proses *peer review*. Cakupan memungkinkan adanya publikasi atau temuan relevan yang belum terakomodasi dalam sintesis. Perkembangan *generative AI* yang berlangsung secara cepat juga menyebabkan karakteristik model, praktik *deployment*, dan kerangka regulasi terus mengalami perubahan. Temuan yang disajikan perlu dipahami sebagai representasi konseptual berdasarkan tingkat kematangan evidensi pada periode yang dikaji, bukan sebagai evaluasi klinis yang bersifat final.

Daftar Pustaka

- AlMakinah, R., Norcini-Pala, A., Disney, L., & Canbaz, M. A. (2024). *Enhancing Mental Health Support through Human-AI Collaboration: Toward Secure and Empathetic AI-enabled chatbots*.
- Attanasio, F., Fazio, V., Antonini, C., Lanzano, N., Obumselu, G., Prato, M., Flutti, E., Pacchioni, F., Fregna, L., Franchini, L. A. M., & Colombo, C. (2026). Evaluating ChatGPT-generated psychoeducation for mood disorders: comparative insights from patients and mental health professionals. *Journal of Psychiatric Research*, 193, 354–360. <https://doi.org/10.1016/j.jpsychires.2025.11.039>
- Blease, C., & Rodman, A. (2024). Generative Artificial Intelligence in Mental Healthcare: An Ethical Evaluation. *Current Treatment Options in Psychiatry*, 12(1), 5. <https://doi.org/10.1007/s40501-024-00340-x>
- Blease, C., & Torous, J. (2023). ChatGPT and mental healthcare: balancing benefits with risks of harms. *BMJ Mental Health*, 26(1), e300884. <https://doi.org/10.1136/bmjment-2023-300884>
- Boit, S., & Patil, R. (2025). A Prompt Engineering Framework for Large Language Model-Based Mental Health Chatbots: Conceptual Framework. *JMIR Mental Health*, 12. <https://doi.org/10.2196/75078>
- Cho, H. N., Zheng, K., Wang, J., & Hu, D. (2025). Large Language Model-based Chatbots and Agentic AI for Mental Health Counseling: A Systematic Review of Methodologies, Evaluation Frameworks, and Ethical Safeguards (Preprint). *JMIR AI*. <https://doi.org/10.2196/80348>
- Christiawan, J., Setiawan, D. A., Oroh, J. A. P., & Hidayaturrehman. (2026). *Sanity-LLM: Enhancing Mental Health Chatbots with Retrieval-Augmented Generation (RAG)*. 1020–1025. <https://doi.org/10.1109/icimcis68501.2025.11327020>
- Collins, A. C., Lekkas, D., Heinz, M. V., Annor, J., Ruan, F., & Jacobson, N. C. (2025). ChatGPT as therapy: A qualitative and network-based thematic profiling of shared experiences, attitudes, and beliefs on Reddit. *Journal of Psychiatric Research*, 191, 277–284. <https://doi.org/10.1016/j.jpsychires.2025.09.057>
- Elboardy, A. T., Mohamed, Z., Khoriba, G., Arafa, T., & Rashed, E. A. (2025). Bridging Gender Disparities in Mental Health: A Bilingual Large Language Model for Multilingual Therapeutic Chatbots. *22nd International Learning and Technology Conference: Human-Machine Dynamics Fueling a Sustainable Future, L and T 2025*, 251–256. <https://doi.org/10.1109/LT64002.2025.10941308>

- Elyoseph, Z., & Levkovich, I. (2024). Comparing the Perspectives of Generative AI, Mental Health Experts, and the General Public on Schizophrenia Recovery: Case Vignette Study. *JMIR Mental Health*, 11(1). <https://doi.org/10.2196/53043>
- Fanarioti, A. K., & Karpouzis, K. (2025). Artificial Intelligence and the Future of Mental Health in a Digitally Transformed World. *Computers*, 14(7), 259. <https://doi.org/10.3390/computers14070259>
- Golden, A., & Aboujaoude, E. (2024). Describing the Framework for AI Tool Assessment in Mental Health and Applying It to a Generative AI Obsessive-Compulsive Disorder Platform: Tutorial. In *JMIR Formative Research* (Vol. 8). JMIR Publications Inc. <https://doi.org/10.2196/62963>
- Grodiewicz, J. P., & Hohol, M. (2023). Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry*, 14. <https://doi.org/10.3389/fpsy.2023.1190084>
- Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health*, 11, e57400. <https://doi.org/10.2196/57400>
- Hadar-Shoval, D., Asraf, K., Mizrachi, Y., Haber, Y., & Elyoseph, Z. (2024). Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values. *JMIR Mental Health*, 11(1). <https://doi.org/10.2196/55988>
- Held, P., Pridgen, S., Chen, Y., Akhtar, Z., Amin, D., & Pohorence, S. (2024). A Novel Cognitive Behavioral Therapy–Based Generative AI Tool (Socrates 2.0) to Facilitate Socratic Dialogue: Protocol for a Mixed Methods Feasibility Study. *JMIR Research Protocols*, 13. <https://doi.org/10.2196/58195>
- Huang, Y., Wang, W., Zhou, J., Zhang, L., Lin, J., Liu, H., Hu, X., Zhou, Z., & Dong, W. (2025). Integrative modeling enables ChatGPT to achieve average level of human counselors performance in mental health Q&A. *Information Processing and Management*, 62(5). <https://doi.org/10.1016/j.ipm.2025.104152>
- Jesudason, D., Bacchi, S., & Bastiampillai, T. (2025). Artificial intelligence (AI) in psychotherapy: A challenging frontier. *Australasian Psychiatry*, 33(4), 629–632. <https://doi.org/10.1177/10398562251346075>
- Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *Npj Digital Medicine*, 6(1), 236. <https://doi.org/10.1038/s41746-023-00979-5>
- Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, 3, 111–132. <https://doi.org/10.1016/j.aiopen.2022.10.001>
- Lozoya, D. C., Conway, M., De Duro, E. S., & D'Alfonso, S. (2025). Leveraging Large Language Models for Simulated Psychotherapy Client Interactions: Development and Usability Study of Client101. *JMIR Medical Education*, 11. <https://doi.org/10.2196/68056>
- Luo, X., Wang, Z., Tilley, J. L., Balarajan, S., Bassey, U. A., & Cheang, C. I. (2025). Seeking Emotional and Mental Health Support From Generative AI: Mixed-Methods Study of ChatGPT User Experiences. *JMIR Mental Health*, 12. <https://doi.org/10.2196/77951>
- Na, H., Hua, Y., Wang, Z., Shen, T., Yu, B., Wang, L., Wang, W., Torous, J., & Chen, L. (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>
- Nesse, R. M. (2023). Evolutionary psychiatry: foundations, progress and challenges. *World Psychiatry*, 22(2), 177–202. <https://doi.org/10.1002/wps.21072>
- Omar, M., Soffer, S., Charney, A. W., Landi, I., Nadkarni, G. N., & Klang, E. (2024). Applications of large language models in psychiatry: a systematic review. *Frontiers in Psychiatry*, 15. <https://doi.org/10.3389/fpsy.2024.1422807>
- Schnepper, R., Roemmel, N., Schaefert, R., Lambrecht-Walzinger, L., & Meinschmidt, G. (2025). Exploring Biases of Large Language Models in the Field of Mental Health: Comparative Questionnaire Study of the Effect of

- Gender and Sexual Orientation in Anorexia Nervosa and Bulimia Nervosa Case Vignettes. *JMIR Mental Health*, 12. <https://doi.org/10.2196/57986>
- Scholich, T., Barr, M., Stirman, S. W., & Raj, S. (2025). A Comparison of Responses from Human Therapists and Large Language Model-Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. *JMIR Mental Health*, 12. <https://doi.org/10.2196/69709>
- Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Npj Mental Health Research*, 3(1), 12. <https://doi.org/10.1038/s44184-024-00056-z>
- Surana, K., Sharma, A., Tibrewal, V., & Kushwaha, R. (2025). *GenAI-based Mental Health Chatbots: Efficacy, Personalization and Ethical Considerations*. 1–6. <https://doi.org/10.1109/ic366947.2025.11290665>
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>
- Voultsiou, E., & Moussiades, L. (2026). A Systematic Review of Large Language Models in Mental Health: Opportunities, Challenges, and Future Directions. *Electronics*, 15(3), 524. <https://doi.org/10.3390/electronics15030524>
- Vowels, L. M., Francois-Walcott, R. R. R., & Darwiche, J. (2024). AI in relationship counselling: Evaluating ChatGPT's therapeutic capabilities in providing relationship advice. *Computers in Human Behavior: Artificial Humans*, 2(2), 100078. <https://doi.org/10.1016/j.chbah.2024.100078>
- Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S., & Hightow-Weidman, L. (2025). Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations. *JMIR Mental Health*, 12, e70014–e70014. <https://doi.org/10.2196/70014>
- Yirmiya, K., & Fonagy, P. (2025). Mentalizing Without a Mind: Psychotherapeutic Potential of Generative AI. In *Journal of Medical Internet Research* (Vol. 27). JMIR Publications Inc. <https://doi.org/10.2196/79156>
- Zhang, Q., Zhang, R., Xiong, Y., Sui, Y., Tong, C., & Lin, F.-H. (2025). Generative AI Mental Health Chatbots as Therapeutic Tools: Systematic Review and Meta-Analysis of Their Role in Reducing Mental Health Issues. *Journal of Medical Internet Research*, 27, e78238–e78238. <https://doi.org/10.2196/78238>
- Zhang, R., Meng, H., Neubronner, M., & Lee, Y.-C. (2025). Computational and ethical considerations for using large language models in psychotherapy. *Nature Computational Science*, 5(10), 854–862. <https://doi.org/10.1038/s43588-025-00874-x>