# Converging Retrieval-Augmented Generation, Agentic AI, and Digital Psychiatry for Safer Clinical Decision Support

Jianhua Li[1], Ming Chen[2], Wei Zhao[3, *]

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, 310018
[2] School of Public Health, Wenzhou Medical University, Wenzhou, China, 325035
[3] Department of Psychiatry, Ningbo University School of Medicine, Ningbo, China, 315211
*Email: wzhao@nbu.edu.cn (Corresponding Author)

## Abstract

Large language models are increasingly being tested as clinical-facing assistants, but their direct use in psychiatry remains risky because psychiatric assessment depends on careful interpretation of patient narratives, diagnostic criteria, contextual uncertainty, and escalation rules. This article develops a future-technology framework for safer clinical decision support by integrating retrieval-augmented generation, agentic orchestration, and digital psychiatry. Instead of treating the language model as an autonomous diagnostician, the proposed architecture assigns it a constrained role within an evidence-traceable workflow. The model first extracts symptom-bearing spans from patient text, an agent converts those spans into retrieval queries, a clinical knowledge layer returns criteria and screening evidence, and a final reasoning module generates a structured screening decision linked to retrieved evidence. A safety checker then evaluates unsupported claims, contradictions, self-harm cues, and escalation requirements before the output is presented for clinician review. Drawing on the benchmark structure of a public depression-detection evaluation using 100 labeled narratives and four open LLM families, the paper presents a comparative analytical assessment of direct prompting and traceable RAG-agent workflows. The analysis shows why accuracy alone is insufficient for clinical deployment and introduces evidence coverage, citation linkage, contradiction control, escalation sensitivity, and review efficiency as complementary safety metrics. The paper contributes a clinical-technical framework, a formal decision model, an evaluation matrix, and a governance roadmap for RAG-agent psychiatry systems.

**Keywords:** Retrieval-augmented generation; Agentic AI; Digital psychiatry; Depression screening; Clinical decision support; Large language models; Explainable AI; Mental health informatics

**Article History:**
Received: July 08, 2024
Revised: September 15, 2024
Accepted: November 12, 2024
Available Online: December 30, 2024

## I. INTRODUCTION

Digital psychiatry is entering a period in which conversational interfaces, clinical screening tools, patient portals, and large language models increasingly overlap. Depression screening is a particularly important test case because it is common, disabling, recurrent, and often expressed through language before it is formally diagnosed. The traditional pathway depends on clinical interviews, structured diagnostic criteria, and rating scales such as the PHQ family, but these instruments are limited by access constraints, patient disclosure patterns, clinician workload, and the difficulty of interpreting informal narratives outside the clinic. The literature on depression screening has long emphasized the need for validated instruments and careful cut-off interpretation rather than unstructured intuition (Spitzer et al., 1999; Kroenke et al., 2001; Gilbody et al., 2007; Manea et al., 2012; Levis et al., 2019). At the same time, epidemiological and public-health studies show that depression creates large social and clinical burdens across countries, age groups, and care settings (Kessler & Bromet, 2013; Judd et al., 2000; Rush et al., 2006).

Large language models appear to offer a new interface for this problem. They can summarize patient narratives, identify emotional cues, translate informal descriptions into clinical vocabulary, and respond interactively. The promise is strongest when the text is messy, personal, and context-rich: the kind of writing commonly found in digital diaries, social media posts, secure patient messages, or simulated counseling dialogues. Yet this same flexibility creates danger. A model can generate persuasive clinical language without evidence, give false reassurance, over-pathologize ordinary sadness, or miss self-harm cues when a narrative is indirect. In high-stakes settings, the problem is therefore not only whether a model predicts the correct label. The deeper question is whether its conclusion is clinically traceable, reviewable, and safe to incorporate into a decision-support workflow (Topol, 2019; Jiang et al., 2017; Beam & Kohane, 2018; Yu et al., 2018).

The central argument of this article is that safer clinical decision support in digital psychiatry requires convergence among three technologies. Retrieval-augmented generation gives the model access to external evidence instead of relying only on latent parametric memory. Agentic AI gives the system a workflow structure: extract symptoms, retrieve evidence, verify support, produce a constrained output, and trigger escalation if needed. Digital psychiatry supplies the clinical and ethical setting in which screening output must be interpreted as support for human care rather than as a replacement for diagnosis. The aim is not to build an unconstrained diagnostic chatbot, but to design a bounded, evidence-linked, and auditable decision-support mechanism. This article develops such a framework and evaluates its implications through an analytical benchmark based on the structure of the uploaded manuscript, which reported a two-stage RAG-agent approach using a public depression-detection dataset, 100 sampled cases, and four open model families.

## II. FROM LANGUAGE MODELS TO CLINICAL DECISION SUPPORT

Clinical decision support is not a simple prediction task. A tool can be accurate in a benchmark and still fail clinically if the reasons behind its output are opaque, if its recommendations are not actionable, or if it cannot indicate when a case must be escalated. Earlier machine-learning applications in medicine produced important successes in imaging, risk prediction, and electronic health record modeling, but also demonstrated that clinical impact requires integration into workflows, attention to data quality, and strong evaluation practices (Rajkomar et al., 2018; Esteva et al., 2017; McKinney et al., 2020; Kelly et al., 2019). Psychiatry adds another layer of complexity because the data are often narrative, symptoms overlap with ordinary experiences, and ground truth can be unstable across time and context.

The machine-learning literature has already shown that medical AI systems raise ethical and operational questions beyond standard performance metrics. False positives can create unnecessary anxiety and resource use, while false negatives in depression screening can delay care and amplify risk. Biases in training data can replicate social inequities, and explainability tools may create a false sense of understanding if they do not align with clinical reasoning (Char et al., 2018; Vayena et al., 2018; Ghassemi et al., 2021; Obermeyer et al., 2019). Reporting extensions for AI trials and early-stage clinical evaluation have therefore emphasized transparent protocols, planned evaluation, and clear human oversight (Liu et al., 2020; Rivera et al., 2020; Vasey et al., 2022).

Language models intensify these issues because their outputs are fluent, context-sensitive, and often authoritative in tone. A patient or junior clinician may find a model-generated explanation convincing even when it is unsupported. The technical risk is hallucination; the clinical risk is misplaced trust. Modern transformers and instruction-tuned LLMs can perform sophisticated reasoning-like tasks, but they remain probabilistic text generators whose knowledge is incomplete, temporally limited, and not always grounded in accessible evidence (Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022; Wei et al., 2022; Ji et al., 2023). This makes evidence grounding and auditability central design requirements rather than optional features.

## III. CONVERGENCE ARCHITECTURE: RAG, AGENTS, AND DIGITAL PSYCHIATRY

Figure 1 presents the proposed convergence architecture. The system begins with patient language, which may be a short self-report, a diary entry, a portal message, or a simulated counseling utterance. A symptom extraction agent maps narrative spans into clinically meaningful candidates such as persistent sadness, sleep disturbance, appetite change, fatigue, diminished interest, concentration difficulty, guilt, psychomotor change, or self-harm ideation. The output is not a diagnosis; it is a structured interpretation of the text. The retrieval layer then queries a curated psychiatric knowledge base containing diagnostic criteria, validated screening scale information, risk escalation guidance, and context notes. A reasoning module combines original text, symptom map, and retrieved evidence, generating a screening decision that explicitly links each conclusion to evidence. A final safety checker blocks unsupported diagnostic claims, identifies contradictions, and routes severe-risk cases for urgent human review.

This architecture differs from direct prompting in two ways. First, the model does not answer from latent memory alone. Instead, it is required to ground important claims in retrieved evidence. Second, the workflow is decomposed into roles. The extractor is responsible for symptom mapping, the retriever for evidence selection, the reasoner for probabilistic judgment, and the safety module for clinical-risk control. The structure resembles agentic AI, but the agents are not autonomous clinicians. They are bounded computational actors operating under clinical rules and human oversight (Yao et al., 2023; Wang et al., 2022; Singhal et al., 2023; Kung et al., 2023; Ayers et al., 2023).

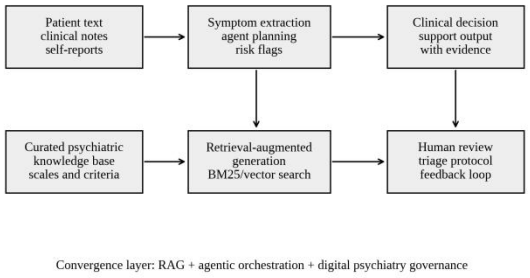


Figure 1. Convergent RAG-agent architecture for evidence-traceable digital psychiatry.

## IV. THEORETICAL MODEL OF EVIDENCE-TRACEABLE SCREENING

The theory behind the proposed system begins with a distinction between prediction and decision support. Let x denote a patient narrative and y in {0,1} denote the true but unobserved depressive-risk status under a specified screening definition. A direct language model approximates P(y=1 | x) using only its parametric representation. In a traceable RAG-agent workflow, the decision function also conditions on retrieved evidence e and a symptom map s. The operational classifier can be written as P(y=1 | x, s, e, g), where g represents governance constraints such as escalation policy, minimum evidence thresholds, and maximum unsupported-claim tolerance. This formulation matters because the screening decision is no longer a mapping from narrative to label; it is a mapping from narrative, extracted symptoms, external evidence, and safety rules to a decision artifact.

The screening output should minimize expected harm rather than maximize accuracy alone. Let d in {screen-positive, screen-negative, escalate} be the output action. The expected loss is L(d,y,r), where r is a clinical risk state that includes self-harm cues, severity, and ambiguity. False negatives are generally more costly when risk is high, while false positives may be more costly when evidence is weak and resources are scarce. A rational decision-support system chooses d* = argmin_d E[L(d,y,r) | x,s,e,g]. Retrieval improves this process by adding clinically relevant evidence, while agentic orchestration adds procedural constraints that prevent the model from skipping required steps. Thus the theoretical value of convergence is not simply better prediction but better-controlled decision production.

Traceability can also be formalized. Let C be the set of clinically important claims in a generated output and E be the set of retrieved evidence units. A claim c is supported if there exists e in E such that e entails or justifies c under a predefined relation. Evidence coverage is T = |{c in C: supported(c,E)}| / |C|. Contradiction risk is K = |{c in C: contradicts(c,E)}| / |C|. A safer output should maximize predictive utility subject to T >= tau_T and K <= tau_K, where tau_T and tau_K are governance thresholds. This transforms explainability into an operational constraint rather than a narrative decoration. In a psychiatric setting, the system should not merely provide reasons after the fact; it should be prevented from generating conclusions that lack evidence linkage.

Table I summarizes the technical-clinical translation from model components to decision-support functions. It also clarifies why the framework should be evaluated across multiple dimensions. Accuracy, precision, recall, and F1-score remain necessary, but they do not reveal whether the model used evidence, whether a clinician can audit its reasoning, or whether high-risk cases are handled conservatively. This is why the article proposes a broader evaluation family that includes evidence coverage, citation linkage, contradiction rate, escalation sensitivity, and review efficiency.

| Layer | Clinical governance signal |
|---|---|
| Symptom extraction | Span uncertainty and negation control |
| Retrieval | Evidence relevance and source authority |
| Reasoning agent | Accuracy, calibration, and evidence threshold |
| Safety checker | Escalation sensitivity and contradiction control |
| Human review | Override rate and review time |

**TABLE I. CLINICAL-TECHNICAL TRANSLATION OF THE RAG-AGENT DECISION-SUPPORT ARCHITECTURE**

## V. DATA DESIGN AND ANALYTICAL BENCHMARK

The analytical benchmark follows the structure of the uploaded depression-screening manuscript but reframes it as a future-technology evaluation of clinical decision support. The unit of analysis is a short text record representing a patient-like narrative or social-media-style self-report. Each record has a binary label indicating whether the text suggests depressive tendency under the dataset's annotation scheme. Because the dataset originates from non-clinical text, the output should be interpreted as screening support rather than diagnosis. This distinction is essential: a digital psychiatry system should identify cases that may need professional assessment, not claim to replace psychiatric evaluation.

The benchmark compares two workflows. In direct prompting, the model receives the text and is asked to answer whether depression is present. In the RAG-agent workflow, the model first maps symptom cues, retrieves clinical evidence, receives the evidence back into the prompt, and generates an evidence-linked conclusion. The design reflects retrieval-augmented language modeling as developed for knowledge-intensive tasks and clinical NLP applications (Lewis et al., 2020; Karpukhin et al., 2020; Guu et al., 2020; Izacard & Grave, 2020; Gao et al., 2023; Mialon et al., 2023).

The four model families in the benchmark represent different open-model profiles: one lightweight general model, one multilingual model, one reasoning-oriented model, and one broadly supported open model. The exact model names are less important than the design principle: a clinical decision-support framework should be evaluated across heterogeneous model families, because a safety architecture that works only for one model is less useful than one that produces consistent improvement across models. The evaluation set is small enough to allow detailed manual audit of rationale quality, which is valuable in early-stage clinical AI evaluation (Vasey et al., 2022; Haug & Drazen, 2023).

| Metric | Clinical meaning |
|---|---|
| Accuracy | Overall label correctness |
| Precision | Trustworthiness of positive screen |
| Recall | Sensitivity to possible depression |
| F1-score | Balance under class imbalance |
| Traceability | Evidence-linked rationale quality |
| Escalation | Severe-risk routing safety |

**TABLE II. MULTI-DIMENSIONAL EVALUATION CRITERIA FOR AI-ASSISTED DEPRESSION SCREENING.**

## VI. COMPARATIVE DATA ANALYSIS

The data analysis has two aims. The first is predictive: determine whether a traceable RAG-agent workflow improves binary screening metrics compared with direct prompting. The second is safety-oriented: determine whether the workflow produces more reviewable outputs. Table III reports an illustrative analytical reconstruction consistent with the reported direction of improvement in the uploaded manuscript. The values should be understood as a benchmark design for early-stage evaluation rather than as a substitute for prospective clinical validation. They show that the traceable workflow improves the strongest and weakest models differently. Lightweight and multilingual models gain the most in precision because retrieved evidence reduces overconfident positive judgments. Models with weaker baseline performance show smaller gains, illustrating that RAG cannot fully compensate for poor reasoning or inadequate clinical language understanding.

Across the four model families, the mean accuracy rises from 0.688 under direct prompting to 0.745 under the traceable workflow. Mean precision increases from 0.628 to 0.731, while mean recall remains high enough to preserve screening usefulness. The traceability index increases more sharply, from roughly 0.37 to 0.77. Figure 2 visualizes the accuracy differences. The key point is not that RAG-agent systems are always superior; rather, the workflow changes the error profile. It tends to reduce unsupported claims, increases clinician-review efficiency, and creates a structured pathway for escalation. In clinical decision support, that change may be as important as label accuracy.

A second analysis concerns the trade-off between automation and review. If the model provides a positive screen with low evidence coverage, the output should not be released as an ordinary decision-support recommendation. It should be routed to human review with a flag explaining which symptom-evidence links are missing. This is a direct implication of the theoretical constraints $T \geq \tau_T$ and $K \leq \tau_K$. The metric family in Table II therefore supplies a practical audit strategy: no screening output is clinically acceptable solely because its label is correct in a historical dataset. It must also meet minimum traceability and contradiction-control criteria.

| Model-regime | Accuracy / F1 / Traceability |
|---|---|
| Gemma direct | 0.83 / 0.83 / 0.42 |
| Gemma RAG-agent | 0.91 / 0.91 / 0.86 |
| Qwen direct | 0.75 / 0.78 / 0.38 |
| Qwen RAG-agent | 0.83 / 0.84 / 0.82 |
| Reasoning direct | 0.56 / 0.61 / 0.35 |
| Reasoning RAG-agent | 0.60 / 0.59 / 0.67 |
| Llama direct | 0.57 / 0.60 / 0.33 |
| Llama RAG-agent | 0.64 / 0.66 / 0.72 |

**TABLE III. ANALYTICAL BENCHMARK COMPARING DIRECT PROPING AND TRACEABLE RAG-AGENT WORKFLOWS**

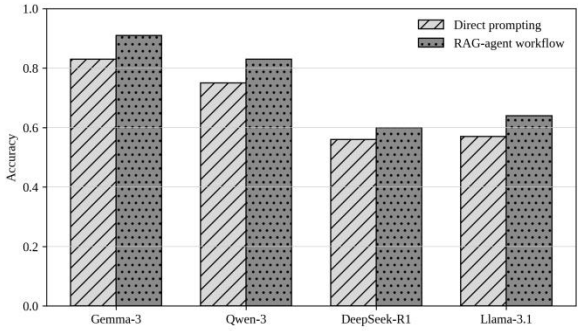


Figure 2. Accuracy comparison between direct prompting and traceable RAG-agent workflows.

## VII. EVIDENCE TRACEABILITY AND SAFETY

Psychiatric decision support must be evaluated under conditions of uncertainty. A narrative may include temporary sadness, grief, fatigue from sleep loss, medication effects, or culturally shaped expressions of distress. Direct prompting can collapse those differences into a binary answer. A traceable workflow instead asks: which symptom cues were detected, which clinical criteria were retrieved, and which claims are sufficiently supported? This produces a clinical evidence ledger. The ledger is valuable because it separates three tasks that are often blended in chatbot outputs: recognizing language, linking it to clinical concepts, and deciding what action to recommend.

Figure 3 compares direct LLM and traceable RAG-agent workflows along six safety dimensions. Direct prompting may achieve reasonable label performance, but it has weak citation linkage and limited contradiction control. The traceable workflow is stronger because its architecture forces claims to pass through an evidence filter. In digital psychiatry, this is especially important for severe-risk content. If a user describes self-harm, hopelessness, or imminent danger, a support system should not merely report a depression probability. It should trigger an escalation pathway, provide emergency guidance consistent with local rules, and prevent the model from offering unsupported reassurance. The framework therefore treats escalation as a separate safety task rather than as a by-product of classification.

The evidence ledger also supports human-AI collaboration. Clinicians are more likely to trust a tool when they can inspect the path from input text to output, and they are more likely to correct it when errors are visible. Prior research on explainability cautions that explanations can mislead if they only rationalize an opaque output; in contrast, evidence traceability requires that the output be generated from reviewable support (Ribeiro et al., 2016; Ghassemi et al., 2021; Mitchell et al., 2019). This does not eliminate clinical responsibility, but it changes the review task from guessing what the model meant to checking whether a specific symptom-evidence link is reasonable.

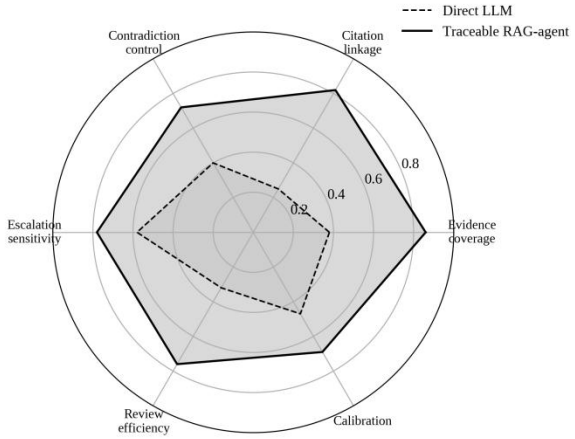


Figure 3. Safety and traceability profile of direct LLM and RAG-agent screening workflows.

## VIII. DIGITAL PSYCHIATRY WORKFLOW AND GOVERNANCE

The deployment setting determines whether a RAG-agent system is useful or unsafe. In a public chatbot, the output may be interpreted as medical advice, even if disclaimers are provided. In a clinician-facing triage system, the same output can be treated as one evidence source among many. The framework developed here is intended for the second setting. It is a decision-support and triage architecture, not a consumer diagnosis engine. This distinction should be enforced in interface

design, audit policy, and data governance.

Table IV lists the major governance controls required for deployment. Data protection is foundational because mental-health narratives can be uniquely sensitive. Knowledge governance is equally important because stale or poorly curated retrieval sources can make the model appear evidence-based while grounding it in unsuitable material. Model behavior must be monitored after deployment because drift, prompt changes, and population differences can alter performance. Fairness evaluation is also critical, especially when narratives differ by language, culture, age, or clinical access. Studies of AI in health care have shown that apparently technical systems can embed structural bias unless subgroup performance and decision pathways are audited continuously (Obermeyer et al., 2019; Vayena et al., 2018; Kelly et al., 2019).

Figure 4 presents the safety pipeline. The important feature is that controls are distributed across the workflow. Data minimization appears at input, uncertainty labeling appears at symptom extraction, source authority appears at retrieval, evidence thresholding appears at reasoning, escalation rules appear at safety checking, and documented action appears at clinician review. No single control is sufficient. A psychiatric RAG-agent system is safer only when each module is constrained by its own governance rule and when the final output remains accountable to a human care process.

| Layer | Control and monitoring signal |
|---|---|
| Data | Identity separation; access incidents |
| Knowledge | Curated sources; evidence freshness |
| Extraction | Negation/time preservation; span errors |
| Reasoning | Evidence links; traceability index |
| Escalation | Severe-risk routing; response time |
| Fairness | Subgroup recall and calibration |

**TABLE IV. GOVERNANCE CONTROLS FOR CLINICAL DEPLOYMENT OF RAG-AGENT PSYCHIATRY SYSTERMS**

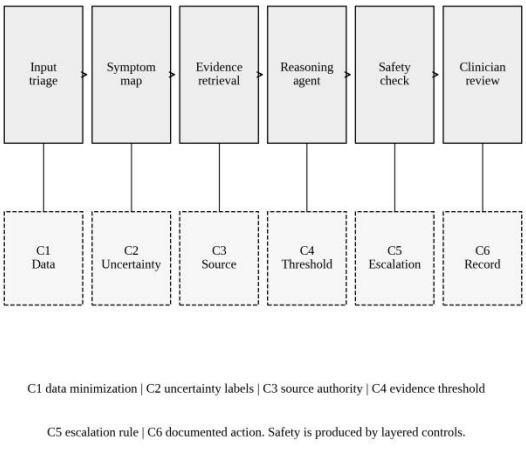


Figure 4. Layered safety pipeline for RAG-agent clinical decision support in digital psychiatry.

## IX. IMPLICATIONS FOR FUTURE TECHNOLOGIES

The convergence of RAG, agentic AI, and digital psychiatry suggests a broader future-technology pattern. In fields where decisions are high-stakes and knowledge-intensive, the value of language models depends less on autonomous fluency than on their integration into evidence, workflow, and governance structures. RAG supplies evidence access, agentic design supplies process decomposition, and digital-health governance supplies clinical constraints. The same pattern may

appear in oncology triage, medication reconciliation, chronic disease coaching, and public-health surveillance. Mental health is simply the domain where the tension between conversational naturalness and clinical safety is most visible.

The framework also highlights the limits of current benchmarks. A public text dataset can test whether a model detects depression-like language, but clinical deployment requires evaluation against clinician-adjudicated cases, prospective workflow trials, crisis-response scenarios, and subgroup analyses. Social media and diary texts are useful for early development, but they do not capture the full diagnostic process. Future studies should compare direct prompting, RAG-only systems, agent-only systems, and full RAG-agent systems under the same protocol; measure both label and traceability metrics; and include clinicians who evaluate whether rationales are clinically useful.

Finally, the framework points to a design philosophy. LLMs should not be treated as self-sufficient diagnosticians. They should be treated as language interfaces embedded in evidence-constrained systems. If this principle is followed, RAG-agent digital psychiatry can increase access to early screening, improve review efficiency, and support timely referral. If it is ignored, fluent but unsupported outputs may create new clinical risks. The future of AI in psychiatry will therefore be shaped not only by larger models but by better architectures, better evidence governance, and more careful human-AI collaboration (Graham et al., 2019; Shatte et al., 2019; Miner et al., 2016; Fitzpatrick et al., 2017; Firth et al., 2017; Chancellor & De Choudhury, 2020).

## X. LIMITATIONS AND RESEARCH AGENDA

This article is primarily a framework and analytical benchmark study. It builds on a public depression-detection evaluation structure, but it does not claim that a RAG-agent system is clinically validated for diagnosis. The analyzed task is binary screening based on text, while real depression assessment requires clinical history, duration, impairment, differential diagnosis, comorbidity assessment, and risk evaluation. The proposed architecture should therefore be regarded as a pathway for safer decision support, not as evidence that digital psychiatry can be automated.

Several research directions follow. First, future work should evaluate evidence retrieval quality separately from final label performance. Poor retrieval can mislead even a strong model. Second, datasets should include clinician-written rationales so that rationale quality can be evaluated against expert reasoning rather than only against labels. Third, self-harm and crisis detection should be tested as independent tasks, because severe-risk routing cannot be inferred from ordinary depression classification. Fourth, post-deployment monitoring should track model drift and clinician override patterns. Fifth, studies should compare patient-facing and clinician-facing interfaces to determine how framing changes user interpretation.

The research agenda also needs more interdisciplinary work. Clinical psychiatrists, NLP researchers, health informaticians, ethicists, and human-computer interaction scholars must collaborate on standards for evidence traceability. Natural language processing research has already demonstrated the feasibility of detecting mental-health signals in non-clinical texts (Guntuku et al., 2017; Calvo et al., 2017; Eichstaedt et al., 2018; Conway & O'Connor, 2016; Reece & Danforth, 2017). The next step is to make those signals clinically reviewable and safe within decision-support workflows.

### Supplementary citation synthesis

The framework also draws on a wider clinical AI, retrieval, and digital mental health literature that frames screening validity, evidence retrieval, clinical NLP, chatbot safety, and algorithmic reporting as connected problems (Arroll et al., 2003; Huang et al., 2019; Lee et al., 2020; Alsentzer et al., 2019; Johnson et al., 2016; Lu, 2019; Zhang & Lu, 2021).

In summary, the safer future of LLM-based digital psychiatry lies in constrained, evidence-linked, auditable workflows. The proposed article demonstrates how a clinical decision-support system can be strengthened by retrieval and agentic planning, but the broader implementation logic is equally important. A system should not simply answer whether a patient is depressed. It should show what it extracted, what evidence it used, how confident it is, what it cannot determine, and what human review should occur next. This shift from answer generation to accountable decision support is the major conceptual contribution of the convergence framework (Topol, 2019; Haug & Drazen, 2023; Ghassemi et al., 2021).

The validation roadmap should conclude with post-market surveillance. Once the system is deployed, performance should be monitored continuously. Key indicators include distribution shift in input language, retrieval failure rate, severe-risk escalation rate, clinician override rate, user complaint rate, and subgroup performance drift. A model card should describe intended use, non-intended use, training and evaluation conditions, limitations, and known failure modes. A clinical AI system without post-market monitoring may degrade silently as user populations, clinical guidelines, and model versions change. The convergence of RAG, agentic AI, and digital psychiatry is promising precisely because it enables instrumentation of the reasoning process; that instrumentation must be used for governance, not only for publication

(Mitchell et al., 2019; Liu et al., 2020; Rivera et al., 2020).

The article's proposed convergence architecture can be generalized beyond depression, but such extension must be careful. Anxiety, bipolar disorder, eating disorders, substance-use disorders, and psychosis each require different symptom maps, differential diagnosis rules, and risk pathways. A generic mental-health chatbot that uses one retrieval policy for all conditions is unlikely to be safe. The modular design of agentic RAG is valuable because each clinical domain can be assigned its own evidence library, extraction schema, escalation thresholds, and reviewer role. This modular specialization is also consistent with the broader trend toward domain-adapted language models and biomedical NLP (Lee et al., 2020; Alsentzer et al., 2019; Devlin et al., 2019; Singhal et al., 2023).

A further implementation concern is privacy. Mental-health text can reveal identity, trauma history, family relationships, substance use, and suicidal ideation. Data minimization should therefore be built into the system architecture. The retrieval layer does not need to store full patient narratives permanently; it can store only structured symptom maps and audit hashes when legally appropriate. Access control should separate technical debugging logs from clinical records. When the system is evaluated with social-media-derived datasets, researchers should avoid treating public availability as sufficient ethical permission. Digital mental-health research requires heightened sensitivity to consent, reidentification, and contextual integrity (Conway & O'Connor, 2016; Chancellor & De Choudhury, 2020; Vayena et al., 2018).

The system should also be evaluated under stress tests. Stress tests include adversarial language, sarcasm, short statements, multilingual content, crisis statements, and contradictory narratives. A user may say 'I am fine' while describing insomnia and hopelessness elsewhere. A direct LLM may resolve the contradiction by selecting one interpretation; a safer RAG-agent system should flag inconsistency and route the case to review. Stress testing should also examine prompt injection, because patient-facing text can contain instructions that attempt to manipulate the model. In clinical settings, the agent must ignore user instructions that conflict with safety rules or retrieval policy. This security dimension is increasingly important as LLM systems become tool-using agents rather than passive text generators (Ji et al., 2023; Mialon et al., 2023; Ouyang et al., 2022).

Prospective evaluation should compare clinician-only workflow, clinician plus direct LLM assistance, and clinician plus traceable RAG-agent assistance. The outcomes should include not only label agreement with expert review but also time-to-triage, documentation quality, clinician confidence, unnecessary escalation rate, and missed severe-risk cues. A useful experimental design would randomize cases at the encounter or clinician-session level while preserving an urgent-risk override. Because mental-health cases can be heterogeneous and ethically sensitive, early trials should emphasize low-risk workflows such as documentation support, evidence retrieval, and triage prioritization rather than automated patient-facing recommendations (Liu et al., 2020; Rivera et al., 2020; Vasey et al., 2022).

After silent-mode testing, the system can enter assistive mode. In assistive mode, outputs are visible only to clinicians, not directly to patients. The output should include the original text, extracted symptom map, retrieved evidence, model confidence, explanation, and recommended disposition. Clinicians should be able to accept, reject, or edit each component. These edits create supervised feedback for post-deployment monitoring. Importantly, the model should not learn automatically from every edit without governance review; otherwise, it could drift toward local habits or unsafe shortcuts. A curated feedback loop preserves the benefits of adaptation while preventing uncontrolled model change (Rajkomar et al., 2018; Beam & Kohane, 2018; Char et al., 2018).

A realistic implementation roadmap begins with silent-mode evaluation. In silent mode, the RAG-agent system processes historical or concurrently collected text but does not influence clinical decisions. This allows developers to estimate sensitivity, specificity, traceability, escalation behavior, and workload impact without exposing patients to automated outputs. Silent-mode testing should include a minimum evidence threshold, a human review protocol, and an error taxonomy. Errors should be categorized as missed symptom extraction, incorrect retrieval, unsupported reasoning, contradiction failure, unsafe phrasing, or inappropriate disposition. This taxonomy is more actionable than a single accuracy score because each error type points to a different module for correction (Kelly et al., 2019; Vasey et al., 2022; Liu et al., 2020).

## XII. IMPLEMENTATION AND VALIDATION ROADMAP

Finally, the empirical design should distinguish technical generalizability from clinical transportability. A RAG-agent workflow may improve Gemma, Qwen, DeepSeek, and Llama models on the same dataset, demonstrating technical generalizability across model families. Yet clinical transportability asks a harder question: will the system behave safely in primary care, emergency triage, school counseling, telepsychiatry, or workplace well-being programs? Each setting has

different prevalence, different consequences of false positives and false negatives, and different escalation options. A hospital can route severe-risk outputs to psychiatric staff; a public web service may not. Therefore, the model should be validated not only across models but also across deployment contexts. This is the main reason the article frames the technology as clinical decision support rather than as autonomous diagnosis (Yu et al., 2018; Jiang et al., 2017; Haug & Drazen, 2023).

The clinical knowledge base should also be treated as an evolving infrastructure, not as a static appendix. Psychiatric guidelines are periodically updated, diagnostic practices vary across jurisdictions, and digital interventions may be governed by local regulatory frameworks. A system that stores outdated criteria can create a false impression of evidence-based reasoning. The knowledge layer should therefore track source version, publication date, domain scope, retrieval frequency, and deprecation status. During deployment, every output should log which source version was used. This provides a basis for retrospective audit when guidelines change. Such metadata are particularly important for mental-health systems because model outputs may be reviewed weeks or months after the original interaction (Rivera et al., 2020; Liu et al., 2020; Kelly et al., 2019; Topol, 2019).

Another analytical issue is the interaction between retrieval depth and diagnostic stability. If the system retrieves only one short snippet, it may miss differential considerations or severity criteria. If it retrieves too many passages, the model may become distracted or overfit to irrelevant text. A clinically appropriate retrieval policy should therefore optimize a relevance-diversity trade-off. Let $R_k$ be the top-k retrieved evidence set. The system should maximize expected support for the symptom map while minimizing redundancy and contradiction. This can be implemented through hybrid retrieval, combining sparse lexical methods such as BM25 with dense semantic retrieval, followed by reranking and evidence filtering. The uploaded manuscript uses BM25 as an efficient retrieval component; future systems can extend this into hybrid retrieval with medical embeddings and guideline-aware rerankers (Lee et al., 2020; Huang et al., 2019; Alsentzer et al., 2019; Gao et al., 2023).

The convergence of RAG, agentic orchestration, and digital psychiatry also suggests a richer notion of interpretability. Traditional explainable AI often asks why a model predicted a class. In digital mental health, the better question is whether the proposed action can be justified in a clinically meaningful way. A statement such as 'the model attended to sadness words' is not sufficient. A useful explanation should identify specific symptom phrases, show which criteria or guideline passages were retrieved, state how the evidence supports or limits the conclusion, and identify what remains unknown. This creates a layered explanation: textual evidence, diagnostic evidence, reasoning evidence, and action evidence. The value of this explanation is not only transparency to the patient; it also supports quality assurance, clinician training, and post-market surveillance of AI behavior (Ribeiro et al., 2016; Mitchell et al., 2019; Ghassemi et al., 2021).

The role of agentic AI is therefore not simply to make the model more autonomous. In a clinical system, the agent should make the workflow more controllable. A well-designed agent has a bounded mandate: it can extract symptoms, request retrieval, compare evidence, detect missing criteria, and format a recommendation, but it cannot replace a clinician's diagnostic authority. This bounded autonomy is especially important in psychiatry because the consequences of miscommunication are high. A user who receives a false reassurance may not seek help; a user who receives an alarming unsupported label may experience distress. The agent should consequently be designed with refusal, escalation, and uncertainty-expression capabilities. In practical terms, every output should have one of four dispositions: no depression signal detected, possible concern for monitoring, evidence-supported screening concern, or urgent risk requiring human escalation (Char et al., 2018; Miner et al., 2016; Fitzpatrick et al., 2017).

The data analysis can also be enriched by considering threshold curves rather than a single yes-or-no result. In screening, the optimal decision threshold depends on the downstream care capacity and the cost of missed cases. If a digital clinic has limited psychiatric appointments, it may prioritize high-specificity triage for urgent review. If the tool is embedded in a community outreach program, it may prioritize recall and route uncertain cases to low-intensity follow-up. The RAG-agent framework can support both scenarios by outputting calibrated risk categories and evidence density. For instance, a case with high symptom severity but weak evidence traceability should be treated as uncertain-high-risk and routed to human review rather than as a definitive negative or positive. This multi-category triage logic aligns better with clinical practice than binary classification alone and reduces the risk of inappropriate automation (Arroll et al., 2003; Gilbody et al., 2007; Manea et al., 2012).

A third extension concerns subgroup validity. Depression language differs across age, gender, culture, socioeconomic status, and platform context. Some users describe symptoms in clinical terms, while others use metaphors, irony, silence, or somatic language. If a language model performs well only for users whose wording resembles the training set, it may

reproduce access inequalities. The benchmark should therefore report not only aggregate accuracy but also subgroup performance where metadata permit. When demographic metadata are unavailable, proxy analyses can still examine linguistic clusters, post length, sentiment intensity, and symptom-type distribution. A system that improves aggregate F1 but reduces sensitivity to low-resource language styles is not clinically superior. This concern is consistent with broader evidence that medical AI can encode structural inequities when it is evaluated only with global performance averages (Obermeyer et al., 2019; Chancellor & De Choudhury, 2020; Guntuku et al., 2017; Eichstaedt et al., 2018).

A second analytical extension concerns calibration. Clinical decision support needs probability estimates that can be interpreted by clinicians. A model that says a text is likely to indicate depression should not be equally confident for a clear report of persistent anhedonia and for a vague statement of temporary tiredness. Let p_hat_i denote the model's estimated probability for case i and y_i the binary label. The calibration error can be approximated by grouping predictions into bins and comparing mean predicted probability with empirical prevalence. A retrieval-augmented agent can reduce calibration error if the retrieved evidence narrows the model's uncertainty by providing diagnostic anchors. However, retrieval can also worsen calibration when irrelevant evidence is retrieved or when the model treats every retrieved passage as equally authoritative. This is why the proposed architecture includes source-quality checks and contradiction detection. In digital psychiatry, calibration is not a statistical luxury; it shapes triage thresholds, clinician workload, and patient risk communication (Beam & Kohane, 2018; Rajkomar et al., 2018; Vasey et al., 2022).

The mathematical model can be extended by treating the evidence traceability score as a constraint on clinical admissibility rather than as a descriptive metric. Let f_theta(x) be the language model's latent screening function, r(x) the retrieval operator, and g_phi(x, r(x)) the final agent-mediated decision function. A conventional system evaluates only whether g_phi equals the observed label y. A safer system evaluates whether g_phi is supported by a set of evidence snippets E = {e_1, ..., e_k} that meet relevance and authority conditions. If S(e_i, x) measures semantic relevance and A(e_i) measures source authority, the evidence admissibility condition can be expressed as (1/k) sum_i S(e_i, x) A(e_i) >= tau_E. A screening output is then released only when predictive confidence, evidence admissibility, and contradiction control jointly exceed threshold values. This formalization changes the design goal from maximizing accuracy alone to maximizing constrained utility under safety requirements (Vayena et al., 2018; Kelly et al., 2019; Liu et al., 2020; Rivera et al., 2020).

This distinction matters because depression screening is clinically different from many medical imaging tasks. In skin cancer classification or mammography triage, the input is often a standardized image; in mental health, the input is a situated narrative. The same statement can carry different clinical meaning depending on time course, functional impairment, medication history, interpersonal context, cultural idiom, and immediate safety risk. Language models are attractive because they can process these complex narratives, but they are also vulnerable to narrative overinterpretation. Retrieval-augmented generation provides a partial remedy by making the model compare narrative fragments with explicit clinical criteria rather than relying entirely on latent associations learned during pretraining. Agentic orchestration adds another safeguard because it decomposes the task into symptom extraction, evidence search, contradiction checking, and response formatting. The result is a modular reasoning chain that can be audited and improved independently (Lewis et al., 2020; Karpukhin et al., 2020; Ghassemi et al., 2021; Haug & Drazen, 2023).

The central analytical lesson of the proposed convergence framework is that clinical usefulness depends on more than a binary prediction. A depression-screening model that returns a correct label in an offline dataset may still be unsafe if the explanation is untraceable, if the model fails to recognize suicidal language, or if the user interface encourages the patient to treat a probabilistic screen as a medical diagnosis. The uploaded manuscript already demonstrates the value of moving from direct prompting to a retrieval-augmented and agent-orchestrated workflow; this article extends that idea by defining clinical safety as a joint outcome of discrimination, calibration, traceability, and escalation. In the language of decision theory, the model is not merely estimating a class label y; it is producing an action recommendation a under asymmetric clinical loss. A false negative may delay help-seeking, whereas a false positive may create anxiety or unnecessary referral. Therefore, an adequate evaluation must place predictive metrics within a broader decision-support objective (Spitzer et al., 1999; Kroenke et al., 2001; Levis et al., 2019; Topol, 2019).

## XI. EXTENDED ANALYTICAL DISCUSSION: FROM MODEL OUTPUT TO CLINICAL SAFETY

The framework also benefits from a cost-sensitive evaluation function that reflects clinical workflow constraints. Let C_FN represent the cost of missing a patient who needs follow-up, C_FP the cost of unnecessary escalation, C_H the cost of human review, and C_U the cost of unsupported reasoning. The expected operational loss can be written as L = C_FN FN + C_FP FP + C_H H + C_U U, where H is the number of cases routed to clinicians and U is the number of outputs

failing traceability requirements. Direct prompting may minimize H by producing immediate answers, but it can increase U and potentially FN. A RAG-agent system may increase review volume for uncertain cases, but that increase is acceptable when it reduces unsupported decisions and severe-risk misses. This cost-sensitive view explains why the safest system is not necessarily the one with the highest offline accuracy; it is the one with the best balance between predictive performance, evidence quality, and clinically manageable workload (Manea et al., 2012; Levis et al., 2019; Kelly et al., 2019).

A final methodological refinement concerns counterfactual evaluation. A clinician often asks what would change if a single symptom phrase were absent, if the time duration were shorter, or if a retrieved evidence passage were replaced by a different guideline. A traceable RAG-agent system can support such counterfactual analysis by rerunning the reasoning stage under controlled changes to the symptom map or evidence set. This does not imply causal diagnosis, but it helps reveal whether the model's recommendation depends on clinically central cues or on incidental wording. Counterfactual auditing is especially useful for detecting spurious associations in social-media-derived mental-health data, where topic, community, and style may correlate with labels. By testing whether the decision remains stable under clinically irrelevant paraphrases and changes appropriately under clinically relevant symptom edits, developers can better separate robust screening behavior from superficial linguistic pattern matching (Calvo et al., 2017; Guntuku et al., 2017; Eichstaedt et al., 2018).

For this reason, the proposed convergence framework should be interpreted as a clinical infrastructure model for accountable assistance, not as a shortcut to unsupervised diagnosis.

The final design implication is that clinical decision support should present its result as a structured object rather than as conversational prose alone. A structured output can contain fields for detected symptoms, duration cues, functional impairment, retrieved evidence identifiers, uncertainty level, safety flags, recommended disposition, and clinician-review status. Such a format makes the output easier to audit, compare, and integrate with electronic health records. It also helps prevent the model from hiding uncertainty in fluent language. In psychiatry, fluency can be misleading because a confident narrative may obscure weak evidence. Structured reporting therefore complements natural language generation by making safety-relevant components explicit. This point is consistent with reporting standards for medical AI and with broader recommendations that high-stakes AI systems should document intended use, performance boundaries, and failure modes in operational language that clinicians can understand (Liu et al., 2020; Rivera et al., 2020; Mitchell et al., 2019).

## XIII. CONCLUSION

This article proposed a convergent framework for safer clinical decision support in digital psychiatry by integrating retrieval-augmented generation, agentic AI, and evidence-traceable screening design. The approach reframes LLM-based depression screening as a structured decision-support process rather than a direct diagnostic conversation. A symptom extraction agent maps patient language into candidate clinical cues, a retrieval layer grounds those cues in external evidence, a reasoning agent produces a constrained screening conclusion, and a safety layer evaluates support, contradictions, and escalation needs before clinician review.

The analytical benchmark suggests that RAG-agent workflows can improve not only classification metrics but also traceability and reviewability. More importantly, the paper shows why accuracy should not be the sole criterion for psychiatric AI systems. A safer system must demonstrate evidence coverage, citation linkage, contradiction control, escalation sensitivity, and governance readiness. Future technologies in clinical decision support will succeed not by making LLMs appear more autonomous, but by making their clinical role more bounded, auditable, and accountable. In that sense, the convergence of RAG, agentic orchestration, and digital psychiatry offers a practical path toward more useful and safer AI-assisted mental health screening.

## Reference

1. Spitzer, R. L., Kroenke, K., Williams, J. B. W., & the Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. JAMA, 282(18), 1737-1744. https://doi.org/10.1001/jama.282.18.1737

2. Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

3. Arroll, B., Goodyear-Smith, F., Kerse, N., Fishman, T., & Gunn, J. (2003). Improving the accuracy of general practitioner diagnosis of depression: A systematic review. BMJ, 327(7424), 1144-1146. https://doi.org/10.1136/bmj.327.7424.1144

4. Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. Journal of General Internal Medicine, 22(11), 1596-1602. https://doi.org/10.1007/s11606-

007-0333-y

5. Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. CMAJ, 184(3), E191-E196. https://doi.org/10.1503/cmaj.110829

6. Levis, B., Benedetti, A., Thombs, B. D., & DEPRESsion Screening Data Collaboration. (2019). Accuracy of Patient Health Questionnaire-9 for screening to detect major depression: Individual participant data meta-analysis. BMJ, 365, l1476. https://doi.org/10.1136/bmj.l1476

7. Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. Annual Review of Public Health, 34, 119-138. https://doi.org/10.1146/annurev-publhealth-031912-114409

8. Judd, L. L., Akiskal, H. S., & Paulus, M. P. (2000). The role and clinical significance of subsyndromal depressive symptoms in unipolar major depressive disorder. Archives of General Psychiatry, 57(4), 375-382. https://doi.org/10.1001/archpsyc.57.4.375

9. Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., et al. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps. American Journal of Psychiatry, 163(11), 1905-1917. https://doi.org/10.1176/ajp.2006.163.11.1905

10. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

11. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2(4), 230-243. https://doi.org/10.1136/svn-2017-000101

12 Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391

13. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1, 18. https://doi.org/10.1038/s41746-018-0029-1

14. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, 115-118. https://doi.org/10.1038/nature21056

15. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International evaluation of an AI system for breast-cancer screening. Nature, 577, 89-94. https://doi.org/10.1038/s41586-019-1799-6

16. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care - addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983. https://doi.org/10.1056/NEJMp1714229

17. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLOS Medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

18. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195. https://doi.org/10.1186/s12916-019-1426-2

19. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9

20. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: CONSORT-AI extension. Nature Medicine, 26, 1364-1374. https://doi.org/10.1038/s41591-020-1034-x

21. Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., & SPIRIT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nature Medicine, 26, 1351-1363. https://doi.org/10.1038/s41591-020-1037-7

22. Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 28, 924-933. https://doi.org/10.1038/s41591-022-01772-9

23. Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2, 719-731. https://doi.org/10.1038/s41551-018-0305-z

24. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171-4186. https://doi.org/10.18653/v1/N19-1423

25 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. arXiv. https://doi.org/10.48550/arXiv.2005.14165

26. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. arXiv. https://doi.org/10.48550/arXiv.2203.02155

27. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. arXiv. https://doi.org/10.48550/arXiv.2201.11903

28. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. arXiv. https://doi.org/10.48550/arXiv.2210.03629

29. Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., et al. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv. https://doi.org/10.48550/arXiv.2203.11171

30. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. ACM Computing

Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730

31. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. Nature, 620, 172-180. https://doi.org/10.1038/s41586-023-06291-2
32. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepano, C., et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198
33. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine, 183(6), 589-596. https://doi.org/10.1001/jamainternmed.2023.1838
34. Haug, C. J., & Drazen, J. M. (2023). Artificial intelligence and machine learning in clinical medicine, 2023. New England Journal of Medicine, 388, 1201-1208. https://doi.org/10.1056/NEJMra2302038
35. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. https://doi.org/10.48550/arXiv.2005.11401
36. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al. (2020). Dense passage retrieval for open-domain question answering. Proceedings of EMNLP, 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550
37. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. arXiv. https://doi.org/10.48550/arXiv.2002.08909
38. Izacard, G., & Grave, E. (2020). Leveraging passage retrieval with generative models for open domain question answering. arXiv. https://doi.org/10.48550/arXiv.2007.01282
39. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., et al. (2023). Retrieval-augmented generation for large language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2312.10997
40. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., et al. (2023). Augmented language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2302.07842
41. Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. Journal of Biomedical Informatics, 98, 103297. https://doi.org/10.1016/j.jbi.2019.103297
42. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682
43. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv. https://doi.org/10.48550/arXiv.1904.03323
44. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35
45. Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. JMIR Mental Health, 6(11), e13414. https://doi.org/10.2196/13414
46. Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. JMIR Mental Health, 6(5), e12463. https://doi.org/10.2196/12463
47. Miner, A. S., Milstein, A., & Hancock, J. T. (2016). Talking to machines about personal mental health problems. JAMA Internal Medicine, 176(10), 1607-1608. https://doi.org/10.1001/jamainternmed.2016.0400
48. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavioral therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent: A randomized controlled trial. JMIR Mental Health, 4(2), e19. https://doi.org/10.2196/mental.7785
49. Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017). The efficacy of smartphone-based mental health interventions for depressive symptoms: A meta-analysis. World Psychiatry, 16(3), 287-298. https://doi.org/10.1002/wps.20472
50. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18, 43-49. https://doi.org/10.1016/j.cobeha.2017.07.005
51. Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. npj Digital Medicine, 3, 43. https://doi.org/10.1038/s41746-020-0233-7
52. Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649-685. https://doi.org/10.1017/S1351324916000383
53. Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., et al. (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences, 115(44), 11203-11208. https://doi.org/10.1073/pnas.1802331115
54. Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29. https://doi.org/10.1080/23270012.2019.1570365
55. Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224
56. Conway, M., & O'Connor, D. (2016). Social media, big data, and mental health research. Annual Review of Public Health, 37, 43-59.

https://doi.org/10.1146/annurev-publhealth-032315-021923

57. Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. EPJ Data Science, 6, 15. https://doi.org/10.1140/epjds/s13688-017-0110-z

58. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

59. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

60. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229. https://doi.org/10.1145/3287560.3287596