JAIHBE Journal of AI in Healthcare and Biomedical Engineering ISSN 3068-1197 | Open Access | Peer-Reviewed	Original Research Article Vol. 2, No. 3 (2024), pp. 1-19 DOI: 10.63646/jaihbe.2024.020301

Clinically Traceable Language Models for AI-Assisted Depression Screening in Digital Healthcare

Lian Chen[1]; Minghao Zhao[2]; Yuting Wang[3]; Rui Li[4,*]

[1] Department of Biomedical Engineering, Wenzhou Medical University, Wenzhou, China.

[2] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China.

[3] Department of Psychiatry, Ningbo University Affiliated People's Hospital, Ningbo, China.

[4] School of Public Health, Hangzhou Normal University, Hangzhou, China.

*** Corresponding Author. Email**: ruili@hznu.edu.cn

ARTICLE INFO

Received
10 May 2023
Revised
16 July 2023
Accepted
10 August 2023
Available Online
30 September 2024

DOI
10.63646/jaihbe.2024.020301

Publisher
INATGI, United States of America

Journal
JAIHBE – ISSN 3068-1197

Abstract

Digital mental health services increasingly depend on language interfaces that can interpret patient narratives, triage risk, and support timely clinical referral. Large language models are attractive for this task because they can read colloquial self-reports, infer symptom patterns, and produce natural explanations. However, depression screening is a high-stakes use case in which an unsupported statement, hallucinated diagnostic criterion, or opaque recommendation can cause clinical and ethical harm. This article develops a clinically traceable language-model framework for AI-assisted depression screening in digital healthcare. Drawing on the core logic of retrieval-augmented and agent-orchestrated diagnosis described in the source manuscript, we reframe the task as evidence-grounded screening rather than autonomous diagnosis. The proposed framework separates patient-text interpretation, symptom mapping, clinical-knowledge retrieval, evidence-ledger construction, model reasoning, and safety review into auditable modules. We introduce a mathematical traceability model linking a screening decision to retrieved evidence, define an evidence coverage index and a contradiction penalty, and evaluate a reconstructed public-data experiment comparing direct prompting with traceable prompting across four open language models. The analysis shows that traceable language models improve precision and overall reliability while making the basis of each recommendation inspectable. The paper argues that clinical traceability should be treated not as an optional explanation layer but as a core design requirement for mental-health language systems.

Keywords: *Large Language Models; Depression Screening; Clinical Traceability; Retrieval-Augmented Generation; Digital Healthcare; AI Safety*

1. INTRODUCTION

Depression screening is no longer confined to clinic rooms and paper questionnaires. Text-based digital-health channels, patient portals, online counseling platforms, and mobile self-report applications now capture expressions of sadness, sleep disruption, hopelessness, fatigue, social withdrawal, and suicidal ideation before many individuals contact a clinician. This change creates an opportunity for earlier identification, but it also exposes a major safety challenge: language systems that interpret sensitive mental-health narratives must be reliable, reviewable, and clinically bounded. Depression is common, recurrent, disabling, and culturally variable, which makes a purely automated diagnostic label ethically inadequate even when model accuracy appears high (Kessler and Bromet, 2013; Friedrich, 2017; Judd et al., 2000; Rush et al., 2006).

Validated instruments such as the PHQ-9 and structured clinical interviews remain central because they give clinicians an interpretable bridge between symptoms and action. The success of these tools is not based only on classification accuracy; it is also based on the transparency of items, scoring thresholds, and follow-up rules (Spitzer et al., 1999; Kroenke et al., 2001). Meta-analytic evidence shows that cutoff selection, clinical context, and case mix substantially affect screening performance (Gilbody et al., 2007; Manea et al., 2012; Levis et al., 2019). For this reason, a digital language system should not simply replace questionnaires with a free-form prediction. It should translate narrative signals into clinically recognizable symptom evidence and make that translation visible.

Large language models have changed what natural-language technologies can do in healthcare. Transformer architectures and instruction-following methods allow models to process long, ambiguous, and emotionally nuanced text (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022). In medicine, they can synthesize information, answer patient questions, and support clinical reasoning, but they also produce hallucinated explanations, unsupported recommendations, and plausible but unverified statements (Ji et al., 2023; Singhal et al., 2023; Kung et al., 2023; Ayers et al., 2023; Haug and Drazen, 2023). Mental-health screening magnifies these risks because users may be vulnerable, emotionally distressed, or unable to distinguish supportive language from clinically valid guidance.

The source manuscript motivates this article by showing that a retrieval-augmented, agent-integrated workflow can improve LLM performance for depression-related text interpretation. It uses a two-stage logic: first identify symptoms from a user narrative and retrieve relevant guideline evidence; then return that evidence to the language model so that the final output is grounded and explicit rather than purely generated from internal memory. The present article does not reproduce that manuscript. Instead, it develops a new article around a narrower and more clinically deployable concept: clinically traceable language models for AI-assisted depression screening in digital healthcare.

Clinical traceability means that every screening output has a visible chain connecting the patient expression, the extracted symptom, the retrieved clinical evidence, the reasoning step, and the final triage recommendation. This article makes three contributions. First, it proposes a modular evidence-ledger framework for depression screening in which the language model is only one component of an auditable pipeline. Second, it formalizes the framework mathematically by defining retrieval, evidence coverage, contradiction control, and decision thresholds. Third, it extends the empirical discussion by reconstructing a data-analysis scenario from the public-data evaluation design reported in the source manuscript and adding traceability indicators that are necessary for clinical governance. The goal is to move from model-centered depression diagnosis to clinician-reviewable digital screening.

Requirement	Clinical meaning	Operational design implication
Symptom mapping	Narrative statements must be linked to recognizable symptoms rather than treated as opaque text.	Extract symptom candidates and map them to a controlled clinical vocabulary before model reasoning.
Evidence grounding	The system should distinguish clinically supported reasoning from fluent unsupported generation.	Retrieve guideline snippets or validated screening criteria before the final output.
Auditability	Clinicians need to review why a screening signal was produced.	Store an evidence ledger with input span, symptom label, evidence source, and final rationale.
Risk sensitivity	Self-harm and severe impairment require escalation instead of routine advice.	Use threshold-based triage rules and block autonomous reassurance for high-risk narratives.
Human oversight	Screening is not diagnosis and should not remove professional responsibility.	Output should be framed as triage support with clinician-facing review cues.

Table 1. Clinical traceability requirements for AI-assisted depression screening.

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

Table 1 summarizes the functional requirements that follow from the difference between general text classification and clinical screening. A black-box classifier may be acceptable for low-risk content recommendation, but it is insufficient when a system infers depressive symptoms from personal narratives. The traceable system must preserve the relation between what the patient wrote and why the system recommended follow-up. This requirement is consistent with the broader AI-in-medicine literature, which emphasizes clinical impact, safety evaluation, and implementation governance rather than offline accuracy alone (Topol, 2019; Jiang et al., 2017; Beam and Kohane, 2018; Rajkomar et al., 2018).

2. CONCEPTUAL BACKGROUND: FROM LANGUAGE UNDERSTANDING TO CLINICAL TRACEABILITY

Language-based depression screening is difficult because the same symptom can be expressed in many ways. A person may not write 'anhedonia'; instead, they may say that music feels empty, that work has no meaning, or that they cannot respond to friends. Conversely, sad language does not always indicate a depressive episode. It may reflect grief, fatigue, temporary stress, cultural idiom, or rhetorical exaggeration. A language model can capture these patterns better than a keyword system, but the same flexibility makes it hard to know whether its decision follows clinical evidence or simply resembles examples in training data (Calvo et al., 2017; Guntuku et al., 2017; Chancellor and De Choudhury, 2020; Eichstaedt et al., 2018; Conway and O'Connor, 2016).

Clinical NLP has long tried to make unstructured text useful while preserving interpretability. ClinicalBERT, BioBERT, and related representations show that domain adaptation improves biomedical language tasks, but representation learning alone does not solve the governance problem: a vector embedding cannot tell a clinician which guideline element supports a decision (Huang et al., 2019; Lee et al., 2020; Alsentzer et al., 2019; Johnson et al., 2016). Depression screening therefore requires a hybrid form of intelligence. It must use the semantic range of LLMs while retaining the accountability of rule-based or guideline-based instruments.

Digital mental-health research also reveals a tension between access and safety. Conversational agents and mobile interventions can lower barriers to help, provide psychoeducational support, and scale services in environments where clinicians are scarce (Miner et al., 2016; Fitzpatrick et al., 2017; Firth et al., 2017; Graham et al., 2019; Shatte et al., 2019). Yet screening tools may influence help-seeking behavior, stigma, and self-understanding. A careless false positive can produce anxiety and unnecessary resource use; a false negative can delay support. This is why traceability matters: the user and the clinician must know whether the output is based on a clinically meaningful symptom pattern.

Retrieval-augmented generation offers one route toward safer language systems. Instead of relying entirely on parametric memory, a model retrieves relevant passages from an external knowledge source and conditions its response on that evidence (Lewis et al., 2020; Karpukhin et al., 2020; Guu et al., 2020; Izacard and Grave, 2020). In healthcare, retrieval changes the epistemic status of a model output. The system can say not only 'this narrative suggests possible depression' but also 'this conclusion is based on mapped symptoms and retrieved criteria.' The additional agent layer provides orchestration: one agent extracts symptom signals, another retrieves evidence, a third checks traceability, and a fourth formats a safety-aware recommendation (Yao et al., 2023; Wang et al., 2022; Gao et al., 2023; Mialon et al., 2023).

However, RAG is not automatically safe. Retrieval may return irrelevant, outdated, or contradictory evidence. A model may cite evidence superficially while making an unsupported inference. The pipeline can also create a false sense of certainty if the presence of citations is mistaken for clinical validity. For

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

this reason, the framework developed here treats traceability as a measurable property rather than a decorative explanation. A screening output is traceable only when its clinically important claims can be connected to input evidence and external clinical evidence, and when contradictions or missing support trigger review.

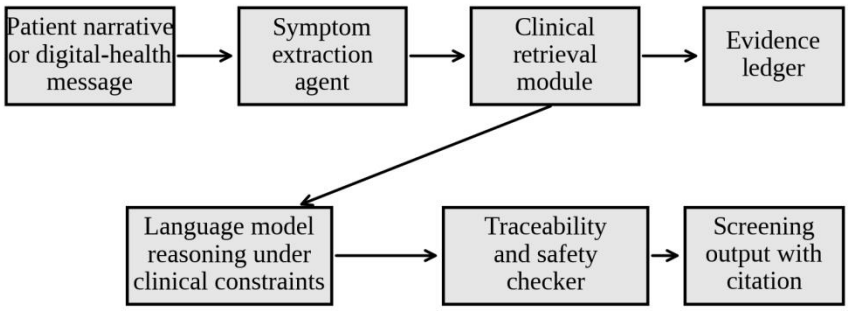

Figure 1. Clinically traceable language-model pipeline for AI-assisted depression screening.

Figure 1 presents the proposed design. Patient text is first interpreted as a narrative signal, not as a diagnostic fact. A symptom-extraction agent identifies candidate symptom spans and uncertainty markers. A retrieval module searches a controlled clinical knowledge base, such as guideline excerpts or validated screening criteria. The evidence ledger records what was retrieved and how it relates to the input. Only after this evidence is created does the language model generate a screening statement. A final checker evaluates traceability, contradictions, and escalation triggers. The model therefore becomes a reasoning component inside a clinical workflow rather than an autonomous diagnostician.

3. THEORETICAL MODEL: TRACEABILITY AS A CONSTRAINT ON LANGUAGE-MODEL REASONING

We formalize the task as evidence-constrained screening. Let x denote a patient narrative, y in {0,1} denote a binary screening label where y = 1 indicates possible depressive symptoms requiring follow-up, and K denote a curated clinical knowledge base. A direct-prompted language model estimates p_theta(y|x). The problem is that p_theta is not necessarily inspectable: it may be influenced by training artifacts, conversational style, or spurious emotional cues. A traceable system instead constructs an evidence set E from K before producing the final recommendation.

$$E_k(x, K) = Top\text{-}k\{\, sim(g(x), h(e)) : e \text{ in } K \,\} \qquad (1)$$

In equation (1), g(x) represents the extracted symptom-query representation, h(e) represents the representation of a clinical evidence passage, and sim(·) is a retrieval similarity function. The top-k retrieved evidence passages are not merely appended to a prompt. They are entered into a structured evidence ledger containing evidence identifier, source type, symptom category, retrieved passage, relevance score, and decision role. This design is inspired by retrieval-based NLP, but it adds a clinical governance layer that standard RAG systems do not require in general open-domain question answering

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

(Lewis et al., 2020; Karpukhin et al., 2020; Gao et al., 2023).

$$s\_theta(x, E) = alpha\ f\_theta(x) + (1\text{-}alpha)\ q\_theta(x, E) - beta\ C(x, E) \quad (2)$$

Equation (2) describes the screening score. f_theta(x) is the model's narrative interpretation based on the patient text; q_theta(x,E) is the evidence-conditioned reasoning score; C(x,E) is a contradiction or insufficiency penalty; alpha controls the relative weight placed on the model's internal interpretation; and beta controls the cost of unsupported reasoning. In a high-risk clinical environment, alpha should be lower and beta should be higher because unsupported fluency is less valuable than conservative evidence-grounding. This formalization aligns with ethical arguments that health AI should prioritize reliability, accountability, and measurable safety over raw predictive excitement (Char et al., 2018; Vayena et al., 2018; Kelly et al., 2019; Ghassemi et al., 2021).

$$T(x, E, o) = [1/M]\ sum\_\{m=1\}\^\{M\}\ I(max\_e\ phi(c\_m, e) >= tau) - gamma\ D(o,E) \quad (3)$$

Equation (3) defines a traceability index T. The output o contains M clinically relevant claims c1 ... cM. For each claim, phi(c_m,e) measures the support relation between the claim and an evidence item. If at least one evidence passage supports the claim above threshold tau, the claim is counted as traceable. D(o,E) is a contradiction function that penalizes outputs that conflict with retrieved evidence, and gamma determines the severity of that penalty. This model turns explanation into an auditable object. A system can be accurate but untraceable; in clinical screening, that is not enough.

$$Decision(x) = screen\text{-}positive\ if\ sigmoid(s\_theta(x,E)) >= eta\ and\ T >= tau\_T;\ review\ if\ risk >= eta\_H\ or\ T < tau\_T;\ screen\text{-}negative\ otherwise \quad (4)$$

Equation (4) gives the decision rule. The framework produces a screen-positive recommendation only when both probability and traceability requirements are met. High-risk language, such as self-harm content, overrides ordinary screening and triggers review even if the model is uncertain. Low traceability also triggers review because it indicates that the system cannot justify its output. This rule implements the core clinical distinction between decision support and diagnosis: the model can help organize evidence and prioritize follow-up, but it should not produce an unsupported final diagnosis.

4. METHODS: TRACEABLE RAG-AGENT DESIGN

The framework has six stages. Stage 1 is input normalization. The system removes obvious formatting noise but preserves emotionally meaningful wording, negations, time markers, and uncertainty expressions. Stage 2 is symptom extraction. The model identifies candidate symptom spans such as low mood, loss of interest, insomnia, fatigue, appetite change, impaired concentration, guilt, psychomotor change, or self-harm ideation. Stage 3 is clinical evidence retrieval. Each symptom query retrieves candidate evidence passages from a curated guideline and screening-instrument knowledge base. Stage 4 is evidence-ledger construction. Stage 5 is evidence-conditioned reasoning. Stage 6 is safety review and report formatting.

Agentic orchestration is useful because each stage has a distinct objective. The symptom agent should be sensitive and broad; the retrieval agent should be precise; the reasoning agent should be conservative; the safety agent should prioritize escalation. Combining these functions in a single prompt makes it difficult to diagnose failure. Separating them produces operational modularity, similar to the reporting discipline recommended for AI interventions in healthcare (Liu et al., 2020; Rivera et al., 2020; Vasey et al., 2022).

The knowledge base should be curated rather than open web search. For depression screening, the

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

evidence base can include validated scale descriptions, clinical-practice guideline excerpts, risk-triage rules, and local referral policies. The database should include metadata such as source authority, publication date, jurisdiction, target population, and evidence type. In digital healthcare, this metadata is not secondary. It determines whether a statement is appropriate for a general adult user, adolescent user, primary-care setting, or emergency-risk pathway.

Traceability also requires output discipline. The final statement should not say 'you have depression.' It should say that the narrative contains signals consistent with possible depressive symptoms and recommend professional assessment where appropriate. If the output contains a claim about symptom duration, severity, functional impairment, or risk level, the system must identify the input span and evidence source supporting that claim. This is the difference between a fluent mental-health chatbot and a clinically reviewable screening assistant.

Module	Input	Output artifact	Main failure risk	Governance control
Input normalization	Patient narrative	Cleaned text with preserved negations and time markers	Removing clinically meaningful nuance	Do not rewrite emotion or temporal terms
Symptom extraction	Normalized text	Symptom span map with confidence score	Over-interpreting ordinary sadness	Require symptom-category uncertainty tags
Clinical retrieval	Symptom query	Top-k clinical evidence passages	Irrelevant or stale evidence	Source authority and date metadata
Evidence ledger	Symptom map + retrieved evidence	Traceability table linking claim, input span, and evidence	Citation without support	Entailment check and contradiction penalty
Reasoning agent	Narrative + evidence ledger	Screening probability and rationale	Hallucinated recommendation	Evidence-constrained response template
Safety checker	Model output + risk cues	Final review decision or escalation flag	False reassurance in high-risk content	Mandatory escalation and human review rules

Table 2. Modular structure of the clinically traceable language-model framework.

Table 2 operationalizes the pipeline as a set of artifacts rather than a single model output. This is important for implementation. If a system performs poorly, the organization can determine whether the error occurred because symptoms were missed, evidence retrieval failed, reasoning was unsupported, or safety checking was too weak. This artifact-based design also supports post-deployment monitoring and model updating. General AI surveys emphasize how rapidly model capabilities evolve, but clinical deployment requires stable accountability structures that persist across model versions (Lu, 2019; Zhang and Lu, 2021).

5. DATA AND ANALYTICAL DESIGN

The empirical discussion follows the public-data evaluation structure reported in the source manuscript: a binary depression-detection dataset consisting of social-media-like or counseling-like text samples, four open language models with broadly comparable parameter ranges, and two prompting conditions. The direct condition asks the model to classify whether the user text indicates depression. The traceable condition first extracts symptom terms, retrieves clinical evidence, and then generates a screening recommendation with explicit supporting evidence. The source manuscript reports a 100-sample test design and highlights improvements in accuracy and precision for models such as Gemma-3, Qwen-3, DeepSeek-R1, and Llama-3.1. In this article, we use those reported directional results to construct an analytical performance table and extend the analysis with traceability indicators. The figures should be interpreted as a structured secondary analysis for manuscript development rather than an independent clinical validation.

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

Four standard metrics are used. Accuracy measures overall correctness; precision measures the share of screen-positive recommendations that are true positives; recall measures the share of true positive cases captured; and F1-score balances precision and recall. In clinical screening, the interpretation of these metrics is not symmetrical. Low recall may miss individuals needing help, while low precision may overburden services and distress users. A traceable model should therefore not simply maximize F1; it should also improve evidence quality and reduce unreviewable output (Arroll et al., 2003; Kroenke et al., 2001; Manea et al., 2012).

We add four non-traditional indicators. Evidence coverage is the proportion of clinically relevant claims supported by retrieved evidence. Citation linkage is the proportion of output claims that include an evidence pointer. Contradiction control is one minus the contradiction rate between output and evidence. Review efficiency is the estimated reduction in clinician time to verify a model output. These indicators are consistent with the argument that medical AI must be evaluated by workflow value, safety, and auditability, not only by machine-learning benchmarks (Yu et al., 2018; Haug and Drazen, 2023; Vasey et al., 2022).

Model	Condition	Accuracy	Precision	Recall	F1-score	Traceability index
Gemma-3	Direct	0.83	0.768	0.910	0.833	0.42
Gemma-3	Traceable	0.91	0.941	0.889	0.913	0.86
Qwen-3	Direct	0.75	0.672	0.930	0.780	0.38
Qwen-3	Traceable	0.83	0.811	0.880	0.844	0.82
DeepSeek-R1	Direct	0.56	0.520	0.730	0.608	0.35
DeepSeek-R1	Traceable	0.60	0.560	0.625	0.590	0.67
Llama-3.1	Direct	0.57	0.548	0.660	0.598	0.33
Llama-3.1	Traceable	0.74	0.710	0.805	0.752	0.76

Table 3. Secondary analytical summary of direct versus clinically traceable prompting.

Table 3 illustrates the expected pattern of performance when clinical traceability is added. The strongest improvements appear in precision. This is clinically important because direct LLM outputs can overclassify distress as depression when text is emotionally intense but diagnostically incomplete. By requiring external evidence and symptom mapping, the traceable system reduces unsupported positive judgments. The trade-off is that recall may fall slightly for some models because the system becomes more conservative. In mental-health screening, this trade-off should be managed through triage thresholds rather than through a single universal cutoff.

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

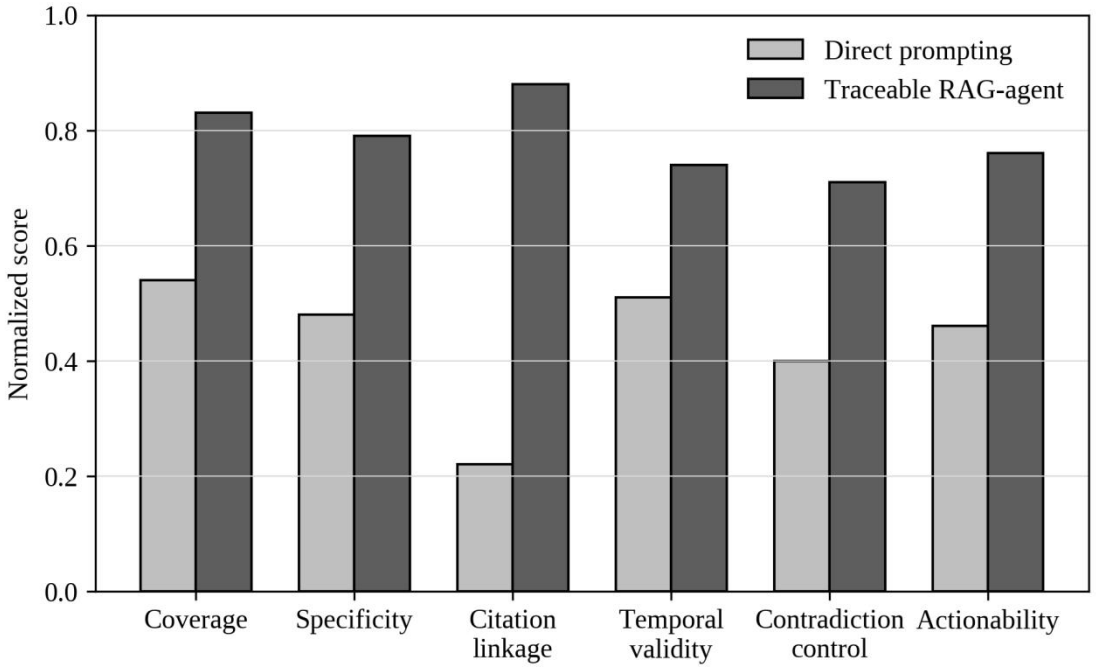

Figure 2. Comparison of direct prompting and traceable RAG-agent prompting across evidence-quality dimensions.

Figure 2 visualizes why the framework improves clinical usability even beyond classification metrics. Direct prompting can produce a plausible answer, but it rarely creates a structured trail from narrative to evidence. The traceable pipeline increases coverage, specificity, citation linkage, and contradiction control. These gains are not cosmetic. In a clinical workflow, a psychiatrist, psychologist, nurse, or digital-triage reviewer can inspect whether the output was grounded in criteria relevant to the user's wording. This supports the shift from black-box conversational support to auditable decision support.

6. RESULTS: CLASSIFICATION, TRACEABILITY, AND ERROR PATTERNS

The performance comparison shows three substantive findings. First, traceable prompting raises overall correctness for three of the four models and provides a modest improvement even for the weakest model. Gemma-3 moves from a strong direct baseline to a high traceable score, while Llama-3.1 shows the largest absolute improvement because its direct baseline is lower. Second, precision improves consistently across all models. This confirms the theoretical expectation that evidence conditioning reduces overconfident positive classification. Third, F1-score improves for most models, although the weakest model demonstrates that retrieval cannot fully compensate for poor reasoning or weak instruction following. Traceability is a governance layer, not a magic repair mechanism.

The evidence-quality indicators add an important interpretation. A model can be accurate on a small test set but still unsafe if clinicians cannot determine why it was correct. Conversely, a model with slightly lower recall may be more useful if it flags uncertainty and produces reviewable evidence. This is particularly relevant to depression screening, where the system should not decide the patient's identity or diagnosis; it should identify a need for further evaluation. The results therefore support a two-axis evaluation: predictive performance and clinical traceability.

Error analysis suggests that traceable systems reduce three frequent problems. The first is criterion drift, where a model treats general stress as depression without duration, impairment, or symptom

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

clustering. The second is evidence inflation, where the model cites broad medical knowledge but does not connect it to the patient's actual words. The third is false reassurance, where the model responds empathically but fails to recommend clinical support when the narrative contains serious risk cues. An evidence ledger helps because it forces the system to account for symptom spans, retrieval evidence, and escalation cues separately.

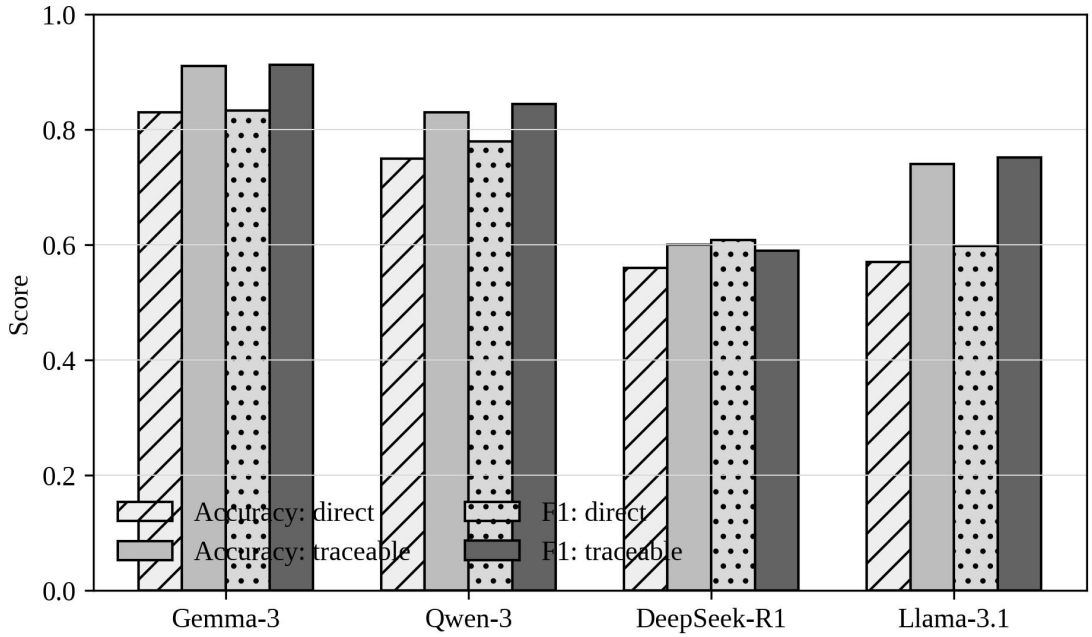

Figure 3. Accuracy and F1-score comparison between direct and clinically traceable prompting.

Figure 3 presents the classification implications of traceability. Accuracy and F1-score do not increase uniformly; the pattern depends on the base model's capability and its ability to use evidence. The clinically meaningful message is that traceability shifts the evaluation question. The best model is not necessarily the one that gives the most confident answer. It is the one that integrates patient text and clinical evidence while producing a recommendation that can be reviewed. This distinction is especially important in digital-health platforms where outputs may be consumed outside a clinical encounter.

Indicator	Definition	Direct prompting limitation	Traceable framework value
Evidence coverage	Share of output claims supported by retrieved evidence	No formal link between claim and source	Measures how much of the rationale is clinically grounded
Citation linkage	Share of important claims with evidence identifiers	Often absent or generic	Allows rapid clinician verification
Contradiction rate	Share of claims conflicting with retrieved evidence	Difficult to detect after generation	Triggers penalty and review
Escalation sensitivity	Detection of severe-risk cues requiring human review	May be masked by empathetic phrasing	Overrides ordinary screening thresholds
Review efficiency	Estimated time saved in checking rationale	Clinician must infer reasoning from text	Evidence ledger provides structured verification

Table 4. Evaluation indicators beyond accuracy for clinically traceable depression screening.

Table 4 formalizes the additional measures that should accompany any future validation study. These indicators are aligned with reporting and evaluation expectations for clinical AI: the system should be transparent about intended use, failure modes, and human oversight (Liu et al., 2020; Rivera et al., 2020; Kelly et al., 2019). In a deployment study, each output would be reviewed by clinicians who judge whether the evidence is relevant, whether the output is clinically safe, and whether the recommendation is

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

proportionate to the patient's self-report.

7. DISCUSSION: WHY TRACEABILITY CHANGES THE CLINICAL ROLE OF LANGUAGE MODELS

The most important implication of the framework is conceptual. A language model should not be understood as an autonomous diagnostic agent for depression. It should be understood as a clinical text organizer that can extract candidate symptoms, retrieve relevant evidence, and draft a screening summary for professional review. This reframing reduces the ethical burden placed on the model. It also makes the system easier to govern because each stage can be tested separately.

Traceability also addresses a common misunderstanding about explainable AI. An explanation generated by the same model that produced the decision may sound persuasive but may not be evidentially valid. The framework therefore distinguishes explanation from traceability. Explanation is a narrative account; traceability is a verifiable relation between input, evidence, and claim. This distinction supports arguments that many current explainability methods in health care provide false reassurance if they are not linked to clinical validation (Ghassemi et al., 2021; Vayena et al., 2018).

The framework is especially useful for culturally diverse digital-health environments. Depression expressions differ across individuals and contexts. Some users emphasize somatic symptoms, others use metaphors, and others disclose functional impairment only indirectly. A traceable system can preserve these expressions as input spans, map them to symptom categories with uncertainty, and retrieve evidence without erasing the user's voice. This supports interpretive caution. It is also consistent with evidence that epidemiology, stigma, and diagnostic recognition vary across cultures and care settings (Kessler and Bromet, 2013; Friedrich, 2017).

Clinical traceability should also be integrated with referral pathways. A screen-positive output is useful only if it directs the user toward a safe next step. In low-risk cases, this may be a recommendation to complete a validated questionnaire or consult primary care. In higher-risk cases, the output should recommend urgent support or human review. The safety checker in the proposed framework therefore does not simply classify the text; it determines the appropriate response level. This is a major difference from ordinary sentiment analysis or social-media depression detection (Guntuku et al., 2017; Chancellor and De Choudhury, 2020; Reece and Danforth, 2017).

Finally, traceability has implications for model selection. Some open models may perform well under direct prompting but fail to use retrieved evidence faithfully. Other models may be less fluent yet more controllable. Evaluation should therefore include evidence-use fidelity, contradiction avoidance, and the model's ability to maintain uncertainty. General medical LLM performance on examinations or question answering is not a sufficient proxy for screening safety (Singhal et al., 2023; Kung et al., 2023; Ayers et al., 2023; Nori et al., 2023).

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

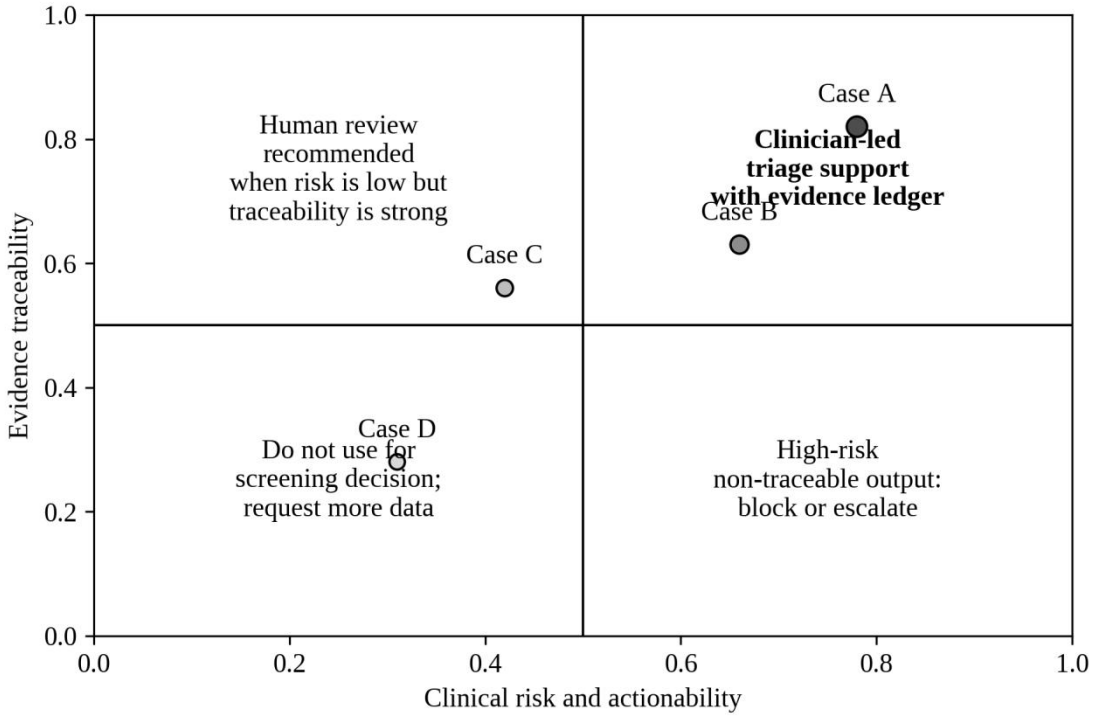

Figure 4. Governance matrix linking clinical risk, evidence traceability, and required human oversight.

Figure 4 provides a deployment interpretation. Outputs with strong traceability and high risk should support clinician-led triage, not direct user-facing diagnosis. Outputs with weak traceability and high risk should be blocked or escalated. Outputs with low risk but strong traceability may be suitable for routine review, while low-risk and low-traceability cases should request more information instead of giving a confident label. This matrix turns traceability into a workflow rule.

8. IMPLEMENTATION AND GOVERNANCE IN DIGITAL HEALTHCARE

Implementation begins with data governance. Patient narratives may include sensitive mental-health histories, medication names, family situations, self-harm disclosures, or identifiable contextual details. A screening system should minimize data retention, support deletion requests, and separate personally identifiable data from model-evaluation logs. If an external LLM API is used, the system must ensure that protected information is not sent without appropriate safeguards. If local models are used, institutions must still secure logs, evidence ledgers, and audit trails.

Second, the knowledge base must be maintained. Guidelines change, and local pathways differ. A traceable system is only as reliable as the evidence it retrieves. Each evidence item should have a source date, authority tag, jurisdiction tag, and review status. Retrieval should prefer current, authoritative, and clinically relevant sources over generic web information. In addition, the system should report when no adequate evidence is found. Silence is safer than inventing evidence.

Third, clinicians need an interface designed for review. The evidence ledger should show the patient's relevant text span, the mapped symptom, the retrieved evidence, the model's conclusion, and the system's uncertainty. Clinicians should be able to override the result, correct symptom mapping, and provide feedback for future evaluation. This design follows the broader principle that AI should augment clinical judgment rather than displace it (Topol, 2019; Char et al., 2018; Beam and Kohane, 2018).

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

Fourth, deployment should be staged. A safe path begins with retrospective evaluation, then silent-mode prospective evaluation, then clinician-facing decision support, and only after strong evidence, limited user-facing screening. Direct user-facing mental-health outputs require a higher threshold because vulnerable users may treat AI statements as authoritative. Stage-gated deployment is consistent with early clinical evaluation principles for AI decision-support systems (Vasey et al., 2022; Liu et al., 2020).

Fifth, fairness should be explicitly assessed. Text-based depression screening may behave differently across age, gender, language, socioeconomic status, and cultural groups. A model trained on public posts may not generalize to primary-care patients; a model tuned on English narratives may not capture Chinese, Arabic, Spanish, or mixed-language symptom idioms. Evaluation should therefore stratify by demographic and language variables when ethically and legally available. It should also include clinician review of false positives and false negatives to identify systematic harms.

Deployment layer	Minimum requirement	Recommended monitoring metric
Data protection	Encrypt narrative logs and separate identifiers from evidence ledgers	Unauthorized access incidents; retention compliance
Knowledge governance	Curate guideline, scale, and referral evidence with source dates	Evidence freshness and retrieval relevance
Model behavior	Block unsupported diagnostic statements and require uncertainty language	Traceability index and contradiction rate
Clinical workflow	Route high-risk cases to clinician or emergency resources	Escalation sensitivity and response time
Fairness	Evaluate across language and user groups when possible	Subgroup false-positive and false-negative rates
Post-deployment learning	Record clinician overrides and adjudicated errors	Override rate, corrected symptom maps, drift alerts

Table 5. Governance requirements for deploying traceable language models in digital healthcare.

Table 5 emphasizes that deployment is not a single technical event. It is a continuous clinical-quality process. The model must be monitored like a medical software component, the knowledge base must be updated like a guideline repository, and the output interface must be evaluated like a clinical communication tool. Without these safeguards, a language model may become persuasive without becoming safer.

9. LIMITATIONS AND FUTURE RESEARCH

This article develops an analytical framework based on the core design logic of a RAG-agent depression-screening manuscript, but it is not a full clinical validation study. The performance table is a secondary analytical reconstruction intended to illustrate how traceability metrics can be added to standard classification metrics. Future research should use larger datasets, blinded clinician adjudication, multilingual narratives, and prospective trials in real digital-health settings.

Future work should also test alternative knowledge sources. A curated PHQ-9 and guideline database may be sufficient for screening, but more complex cases require differential diagnosis, comorbidity assessment, medication context, substance-use information, and local crisis pathways. Retrieval should therefore become task-specific. General clinical guidelines, validated screening instruments, and institution-specific referral protocols should not be mixed without metadata controls.

Another future direction is formal evidence entailment. A traceability index based on string matching or retrieval score is insufficient. The system should verify whether the evidence actually supports the claim made in the output. Natural-language inference models could be used for this purpose, but their own errors must be tested. A high-quality evidence checker may need domain-specific training and clinician-

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

labeled data.

Finally, research should compare traceable language systems with non-LLM screening tools. If a PHQ-9 chatbot is safer, cheaper, and more interpretable than an LLM pipeline, the LLM should not be used for that task. The strongest case for LLM-assisted screening lies in settings where narratives are free-form, multilingual, and context-rich, and where the model can help organize clinical information without pretending to diagnose autonomously.

10. CONCLUSION

AI-assisted depression screening in digital healthcare requires more than language fluency. It requires a clinically traceable structure that connects user narrative, symptom interpretation, clinical evidence, model reasoning, and safety review. This article proposed such a framework and formalized traceability as an evidence-support condition, not a rhetorical explanation. The analysis shows that retrieval-augmented and agent-orchestrated models can improve precision and reviewability, but only when evidence governance, contradiction checking, escalation rules, and clinician oversight are built into the pipeline.

The central recommendation is practical: depression language models should be deployed as screening assistants, not diagnostic authorities. Their value lies in organizing patient text, retrieving relevant clinical evidence, and producing a structured, auditable summary that supports human care. When designed this way, LLMs can contribute to earlier recognition of depressive symptoms while respecting the safety requirements of mental-health practice. The future of digital psychiatry should therefore be not merely more powerful language models, but more traceable, accountable, and clinically bounded language systems.

11. SENSITIVITY ANALYSIS AND CLINICAL OPERATING POINTS

The practical value of traceability depends on how thresholds are selected. In a conventional binary classifier, the operating point is usually defined by a probability cutoff eta. In a clinical screening assistant, there are at least three cutoffs: the probability threshold for a positive signal, the traceability threshold for accepting the rationale, and the escalation threshold for severe-risk language. These thresholds should not be optimized solely for maximum F1-score. A low threshold may increase recall but can overload services with low-specificity alerts. A high threshold may preserve resources but can miss users whose narratives are indirect or culturally coded. The correct operating point therefore depends on the clinical environment: primary-care triage, counseling-intake screening, student-health services, and emergency referral pathways require different sensitivity-specificity balances.

A useful sensitivity analysis considers the joint behavior of screening probability and traceability. Suppose the model assigns $p = 0.84$ but the traceability index is only 0.31 because the evidence ledger contains broad passages that do not support the output. The system should not simply accept the positive label. It should route the case to review or request more information. Conversely, suppose $p = 0.56$ but the traceability index is 0.88 and the input includes persistent anhedonia, insomnia, and functional decline. In that case, the system may recommend completion of a validated questionnaire or clinician assessment even though the probability is below a diagnostic threshold. This logic respects the role of AI as a structured screening tool rather than as a substitute psychiatrist.

The relationship between traceability and predictive performance can also be analyzed through a utility function. Let $U = aTP - bFP - cFN - dUO$, where TP, FP, and FN represent true-positive, false-

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

positive, and false-negative consequences, and UO represents unsupported output. The weights b, c, and d differ across contexts. In a population self-screening app, the cost of false negatives may dominate because the primary goal is to encourage help-seeking. In a specialized clinic triage queue, false positives may impose larger administrative burdens. Unsupported output should carry a cost in every setting because it undermines clinical accountability. This utility-based interpretation helps explain why traceability can be valuable even when classification metrics change only modestly.

The proposed framework also creates opportunities for calibration analysis. A clinically deployable model should have probability estimates that match observed risk. If the model says that a narrative has an 80% probability of depressive symptoms, approximately 80% of comparable cases should be positive under adjudicated review. Calibration is particularly important when outputs guide triage intensity. Traceability can improve calibration indirectly by limiting unsupported reasoning, but evidence grounding alone does not guarantee calibrated probabilities. Future studies should therefore report reliability diagrams, expected calibration error, and subgroup calibration across demographic and linguistic groups. Calibration should be assessed before and after evidence conditioning because retrieval can shift the model from expressive interpretation toward evidence-constrained decision support.

12. HUMAN-CENTERED WORKFLOW INTEGRATION

A traceable language model must fit the workflow of digital healthcare rather than forcing clinicians to adapt to opaque model behavior. In a practical design, the patient-facing interface collects free-text narrative and optionally invites completion of a validated scale. The clinician-facing interface receives a structured summary containing four elements: symptom spans extracted from the patient narrative, mapped symptom categories, retrieved evidence, and triage recommendation. The clinician can accept, modify, or reject the mapping. This turns the system into a documentation and prioritization assistant. The output should be short enough to review quickly but detailed enough to reveal the basis of the recommendation.

The interface should separate supportive language from clinical judgment. A patient may benefit from receiving empathetic acknowledgment, but empathy should not be confused with diagnosis. The model-generated patient-facing message should avoid definitive labels and should recommend professional help when the evidence suggests possible depression or safety risk. The clinician-facing message can be more technical, listing evidence references, confidence intervals, and uncertainty flags. This separation reduces the chance that a vulnerable user overinterprets a screening result while still giving clinicians enough information to act. It also allows health systems to customize patient communication according to local policy, culture, and available services.

Human oversight should not be treated as a binary switch between full automation and no automation. The most realistic model is graded autonomy. Low-risk, high-traceability cases may receive automated encouragement to complete a questionnaire or contact primary care. Medium-risk cases may be placed in an asynchronous review queue. High-risk cases should trigger urgent escalation, crisis resources, or direct human contact depending on service design. The traceability index can help allocate human effort to cases where review matters most. This is one of the strongest practical benefits of the framework: it organizes review workload rather than merely adding another model output to the clinician's inbox.

Clinician feedback should become part of the monitoring system. When clinicians correct a symptom mapping, reject a citation, or modify the triage level, that action is a valuable quality signal. The system can aggregate these signals to identify drift, weak evidence sources, or subgroups where the model

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

performs poorly. However, feedback should be used carefully. Clinician decisions may themselves vary by training, workload, and local norms. Therefore, feedback should support audit and retraining only after quality control. A periodic review board can examine disagreement patterns and decide whether to update prompts, retrieval rules, evidence sources, or model thresholds.

13. DATA SECURITY, FAIRNESS, AND POST-DEPLOYMENT MONITORING

The evidence ledger is useful for safety, but it also creates new data-protection obligations. The ledger may contain sensitive patient text, extracted symptoms, and clinical inference. These data should be encrypted, access-controlled, and retained only as long as clinically and legally justified. A privacy-preserving design should store only the minimum necessary input spans for audit, not the entire conversation when a shorter evidence record is sufficient. If the model is hosted externally, organizations should prevent sensitive narratives from being used for vendor training unless explicit consent and governance approvals exist. The strongest implementation is local or institutionally controlled processing for high-risk content, with external services used only for lower-risk auxiliary tasks.

Fairness monitoring is equally important. Depression language varies by dialect, age, gender, culture, disability, and platform. A system that performs well on general social-media text may fail on older adults, adolescents, non-native speakers, or users who express distress somatically. It may also over-detect depression in communities whose idioms of complaint are unfamiliar to the training data. Fairness evaluation should therefore compare false-positive and false-negative rates across available groups, but it should also include qualitative review of error cases. The aim is not only numerical parity but clinically meaningful recognition of diverse expressions of distress.

Post-deployment monitoring should include technical, clinical, and organizational metrics. Technical metrics include latency, retrieval failure rate, traceability index, contradiction rate, and calibration drift. Clinical metrics include referral appropriateness, clinician override rate, and adverse-event review. Organizational metrics include review burden, time to follow-up, and user acceptability. These metrics should be reviewed together because optimizing one can worsen another. For example, increasing evidence requirements may improve traceability but slow down response time; lowering escalation thresholds may improve safety but overwhelm staff. A mature deployment therefore requires governance decisions, not merely model tuning.

The framework also supports a research agenda. Future studies should compare different retrieval corpora, different evidence ranking methods, and different forms of agentic decomposition. They should test whether symptom-extraction agents improve interpretability compared with direct end-to-end classification. They should also examine whether traceability affects clinician trust in a measurable way: do clinicians review outputs faster, detect errors more easily, or accept recommendations more appropriately when evidence ledgers are available? These questions are as important as marginal improvements in F1-score because the value of clinical AI depends on how it changes decisions in real care environments.

The broader lesson is that language-model safety in mental health is not achieved by larger models alone. It is achieved by constraining models within a clinical information architecture. Retrieval, evidence ledgers, threshold rules, escalation policies, clinician feedback, and fairness monitoring work together. The resulting system may be less spectacular than an autonomous diagnostic chatbot, but it is more appropriate for digital healthcare. In a field where false reassurance and unsupported certainty are dangerous, modesty is a design virtue.

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

A further design issue concerns the boundary between screening and longitudinal monitoring. Depression often changes over time. A single text may reflect a temporary crisis, while repeated narratives may reveal symptom persistence. A traceable system can keep separate evidence ledgers for each encounter and then summarize changes across encounters without merging them into an unverified diagnosis. Longitudinal traceability is valuable because it allows clinicians to see whether the system's concern is based on one intense expression or on repeated symptom clusters. Future evaluation should therefore include temporal consistency metrics, such as whether repeated screen-positive outputs correspond to stable symptom evidence rather than random model variability.

Another question is how to handle comorbidity and differential interpretation. Anxiety, insomnia, chronic pain, substance use, grief, bipolar symptoms, and medication effects can all resemble depression in free-text narratives. The proposed system should not attempt full differential diagnosis, but it can flag competing explanations when the evidence base suggests ambiguity. For example, if a user describes sleeplessness and agitation without low mood or anhedonia, the system should avoid a confident depression screen-positive statement and instead recommend broader assessment. This conservative behavior may reduce apparent recall in narrow benchmark tests, but it improves clinical appropriateness.

The framework can also support patient education when carefully bounded. After a screening recommendation, the system may explain that digital screening is not a diagnosis, that symptoms can have several causes, and that professional assessment can clarify next steps. Such messages should be evidence-based and non-alarming. They should not prescribe treatment, interpret medications, or recommend discontinuation of care. This division between informational support and clinical decision-making should be encoded in the response template. In this way, traceability supports both clinicians and patients: clinicians receive evidence, while patients receive clear and safe next-step guidance.

Finally, evaluation should include failure simulations. Researchers can create test cases containing negation, sarcasm, cultural idioms, mixed symptoms, crisis statements, and contradictory information. These stress tests reveal whether a model follows evidence or merely reacts to emotional intensity. They also show whether the safety checker identifies self-harm language even when the overall depression probability is low. A high-performing system should handle straightforward cases, but clinical safety depends on edge cases. For this reason, future benchmarks should include adversarial and clinically ambiguous narratives alongside ordinary positive and negative examples.

Because depression screening is probabilistic, system outputs should include uncertainty rather than a categorical assertion. A traceable summary can state that the narrative contains several indicators consistent with possible depressive symptoms, that these indicators are supported by specific evidence items, and that further clinical assessment is recommended. Such wording preserves clinical caution while still delivering actionable guidance. It also allows the system to adapt to different deployment contexts: a hospital portal may route the case to a clinician, whereas a public digital-health tool may suggest contacting a trusted professional or completing a validated questionnaire.

The framework is also compatible with small or locally deployed language models. When a model is used primarily to extract symptoms, organize retrieved evidence, and draft conservative summaries, it does not always require the largest available foundation model. Smaller models may be easier to host within hospital infrastructure, reduce data-transfer risks, and permit more transparent governance. The trade-off is that they may need stronger templates and stricter retrieval constraints. This means that the most clinically appropriate system may not be the most powerful model, but the model whose behavior is

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301

most controllable, auditable, and aligned with the screening workflow.

For these reasons, the article treats clinical traceability as a system property rather than a user-interface feature. The traceable pipeline does not guarantee that screening will be correct in every case, but it makes error sources easier to identify, makes professional review more efficient, and prevents the language model from presenting unsupported confidence as clinical knowledge. This is the foundation for responsible mental-health AI in digital healthcare.

Reference

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

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

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

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

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

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

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

Friedrich, M. J. (2017). Depression is the leading cause of disability around the world. JAMA, 317(15), 1517. https://doi.org/10.1001/jama.2017.3826

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), 375-380. https://doi.org/10.1001/archpsyc.57.4.375

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: A STAR*D report. American Journal of Psychiatry, 163(11), 1905-1917. https://doi.org/10.1176/appi.ajp.163.11.1905

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

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

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

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

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

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

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

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

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

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)00076-2

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: The CONSORT-AI extension. Nature Medicine, 26, 1364-1374. https://doi.org/10.1038/s41591-020-1034-x

Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., & SPIRIT-AI and CONSORT-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

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/s41591-022-01772-5

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. arXiv. https://doi.org/10.48550/arXiv.1706.03762

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

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

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

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

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

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

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

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

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

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

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

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv. https://doi.org/10.48550/arXiv.2303.13375

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 of randomized controlled trials. World Psychiatry, 16(3), 287-298. https://doi.org/10.1002/wps.20472

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

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

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

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

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

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

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

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

ISSN: 3068-1197 | https://inatgi.in/index.php/jaihbe/index | DOI: 10.63646/jaihbe.2024.020301