**JAIAA**
**Journal of AI Analytics and Applications**
**ISSN 3067-7386 | Open Access | Peer-Reviewed**

**RESEARCH ARTICLE**
**Vol. 3, No. 3 (2025), pp. 19–35**
**DOI: 10.63646/jaiaa.2025.030302**

# Evidence-Grounded AI Analytics for Explainable Mental Health Screening Using Large Language Models

Wenjun Hao[1], Liping Xu[2], Yuxin Tan[3], Jianhao Zhou[4], Mingxuan Cao[5], Qing Wei[6,*]

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
[2] Department of Psychology, Zhejiang Normal University, Jinhua 321004, China
[3] School of Information Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China
[4] College of Public Health, Wenzhou Medical University, Wenzhou 325035, China
[5] School of Management, Hangzhou Normal University, Hangzhou 311121, China
[6] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
* Corresponding author: Qing Wei, qing.wei@hdu.edu.cn

**ARTICLE INFO**
Received
April 18, 2024
Revised
June 17, 2024
Accepted
August 15, 2025
Available Online
September 30, 2025
DOI
10.63646/jaiaa.2025.030302

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

**Abstract**
Depression is among the most prevalent and disabling mental disorders worldwide, yet timely and accurate diagnosis remains a persistent public-health challenge. Large Language Models (LLMs) have shown promise as auxiliary tools in clinical screening due to their strong language understanding and generation capabilities, but their direct deployment in psychiatric decision support is hindered by hallucination, opacity, and the absence of traceable evidence. To address these limitations, this paper proposes an evidence-grounded analytics framework that integrates Retrieval-Augmented Generation (RAG) with an agent-based two-stage diagnostic pipeline. In the first stage, an LLM agent extracts salient symptom phrases from a user-provided text and formulates a query against a structured knowledge base derived from authoritative clinical practice guidelines. In the second stage, the retrieved evidence is fed back to the LLM, which produces a diagnostic conclusion together with explicit citations to the supporting guideline excerpts. We instantiate the framework on four open-weight LLMs (Gemma-3, Qwen-3, DeepSeek-R1, and Llama-3.1) at the 4–8B parameter scale and evaluate it on a public dataset of 100 simulated counseling samples. The augmented framework increases accuracy by up to 17 percentage points (Llama-3.1: 57% → 74%) and precision by up to 17 percentage points (Gemma-3: 76.81% → 94.12%) compared with direct prompting, while maintaining competitive recall. Two contributions follow: (i) a unified RAG-Agent diagnostic architecture that grounds LLM outputs in verifiable clinical evidence, substantially reducing false positives and improving interpretability; and (ii) a comprehensive empirical study across heterogeneous LLM families demonstrating the cross-model generality of the approach. Our results suggest that evidence-grounded LLM analytics constitute a viable pathway for safe and trustworthy AI deployment in mental-health screening.

**Keywords:** Depression Screening; Large Language Models; Retrieval-Augmented Generation; LLM Agents; Explainable AI; Evidence-Based Reasoning

## 1. Introduction

Depression is one of the most pervasive mental disorders of the modern era, affecting an estimated 280 million

people globally and ranking as a leading contributor to years lived with disability (Liu et al., 2024; Teferra et al., 2024). The disorder manifests as persistent low mood, loss of interest and pleasure, sleep and appetite disturbances, and, in severe cases, suicidal ideation, profoundly disrupting personal, occupational, and social functioning (Wan et al., 2025). Compounding the clinical burden, the conventional diagnostic pathway, which combines structured interviews with psychometric instruments such as the Patient Health Questionnaire-9 (PHQ-9) (Gómez-Gómez et al., 2023; Cumbe et al., 2020), is highly dependent on trained clinicians and remains inaccessible in many primary-care and resource-constrained settings. Recent burden-of-disease analyses indicate that adolescents and young adults are disproportionately affected and that the trend is projected to continue rising (Liu et al., 2025; Erskine et al., 2024). These patterns underscore the urgent need for scalable, low-cost screening technologies that complement, rather than replace, professional assessment.

Recent advances in Large Language Models (LLMs) such as GPT-4, the Llama family, and emerging open-weight models have produced unprecedented capabilities in language comprehension, dialogue, and reasoning (Khan, 2025; Aydin et al., 2025). In mental health, several investigations have examined LLMs as data-driven assistants for risk identification, severity grading, and decision support (Omar et al., 2024; Patel et al., 2024). Promising results have been reported on tasks ranging from depression detection in social-media text (Lan et al., 2024; Kim et al., 2025) to triage of psychiatric emergencies (Chowdhury et al., 2025). Despite this momentum, three challenges have repeatedly limited clinical translation. First, LLMs frequently exhibit hallucination, producing fluent but factually incorrect statements (Pal et al., 2024; Asgari et al., 2025; Chen et al., 2025). Second, their reasoning processes are opaque, leaving clinicians unable to verify why a particular conclusion was reached (Chaddad et al., 2023; Singh et al., 2025). Third, the static nature of their parametric knowledge means they cannot reliably reflect updated clinical guidelines or local practice norms (Ke et al., 2025; Mendoza et al., 2025).

Retrieval-Augmented Generation (RAG) addresses these limitations by coupling the generative power of LLMs with explicit, retrievable evidence drawn from authoritative external sources (Mahmood et al., 2025; Garza et al., 2025). Concurrently, agent-based architectures decompose complex tasks into discrete, controllable sub-tasks that can be orchestrated under structured prompts (Guo et al., 2024; Cao et al., 2024; Tran et al., 2025). The combination of RAG and agent design has emerged as a particularly attractive paradigm for high-stakes clinical applications, where reliability and transparency are paramount (Shi et al., 2024; Wang et al., 2024).

Building on this foundation, this paper introduces an evidence-grounded analytics framework for explainable depression screening that synthesizes RAG with a two-stage agent pipeline. The framework deliberately separates symptom identification from diagnostic reasoning. In the first stage, an LLM-driven agent reads the user's free-form text and extracts a list of clinically salient symptom phrases. These phrases form a query that is dispatched to a curated knowledge base derived from American Psychiatric Association (APA) clinical practice guidelines (Arbanas, 2015) and indexed using the BM25 ranking function (Robertson & Zaragoza, 2009; Li et al., 2024). The second stage feeds the retrieved guideline excerpts back to the LLM, instructing it to map the user's narrative onto the diagnostic criteria and to emit a conclusion together with explicit citations of the supporting excerpts. This deliberate evidence anchor transforms an open-ended generation task into a structured reasoning task whose conclusions are verifiable.

To verify the proposal, we conducted controlled experiments on a public dataset of 100 simulated counseling sessions using four open-weight LLMs (Gemma-3, Qwen-3, DeepSeek-R1, and Llama-3.1) at comparable parameter

scales (Aydin et al., 2025; Khan, 2025). Across these heterogeneous architectures, the augmented framework achieves substantial gains over direct-prompting baselines, with accuracy improvements of up to 17 percentage points and precision improvements of up to 17 percentage points. The contributions of this work are threefold: (1) a unified RAG-Agent framework specifically tailored to evidence-grounded depression screening; (2) a quantitative evaluation that demonstrates cross-model generality; and (3) a discussion of design trade-offs, with particular attention to the precision–recall balance and the governance implications of citation-based outputs in mental-health applications.

The remainder of the paper is organized as follows. Section 2 reviews related work in LLM-based mental-health analytics, RAG, and agent frameworks. Section 3 presents the proposed framework, including the knowledge-base construction, the BM25 retrieval module, and the prompt design. Section 4 details the experimental setup and reports the empirical results. Section 5 analyses the findings and discusses limitations. Section 6 concludes.

Beyond the immediate technical contribution, we believe that the broader value of this study lies in re-framing the role of LLMs in mental-health practice. Rather than positioning the model as an autonomous diagnostician, we argue for a workflow in which the model is treated as an evidence-aware analyst: capable of surfacing relevant clinical knowledge, mapping user narratives onto formal criteria, and producing arguments that practising clinicians can review and adjudicate. This perspective is consistent with the growing consensus that AI systems in healthcare should augment rather than replace human judgement (Thirunavukarasu, 2024; Goh et al., 2024), and aligns with calls for transparent, accountable AI from regulators, professional bodies, and patient-advocacy groups. By demonstrating that this vision is technically achievable using off-the-shelf open-weight models and a modest knowledge base, the present work offers a concrete blueprint that other research groups, hospitals, and digital-health vendors can adapt to their local contexts.

## 2. Related Work

### 2.1 LLMs in Mental Health Analytics

The application of LLMs in mental health has expanded rapidly. Omar et al. (2024) systematically reviewed 16 studies that explored ChatGPT and GPT-4 in psychiatric tasks ranging from clinical reasoning to social-media-based screening, reporting promising but uneven performance. Lan et al. (2024) introduced DORIS, a hybrid system that combines LLM-generated symptom annotations with classical classifiers, achieving improved AUPRC on benchmark depression datasets. Building on transformer-based representations, Wan et al. (2025) proposed a RoBERTa–GRU hybrid for college-student depression detection, while Kim et al. (2025) demonstrated that LLM-derived embeddings substantially boost interpretability when combined with linguistic-feature lexicons. Despite these advances, recent reviews caution that LLM outputs in mental-health contexts can diverge meaningfully from clinical recommendations, especially for complex or comorbid presentations (Patel et al., 2024; Hadar-Shoval et al., 2024). Levkovich and Elyoseph (2024) further showed that ChatGPT differed from primary-care physicians in handling demographic biases, suggesting that uncritical reliance on parametric knowledge is risky. These findings collectively motivate frameworks that constrain LLM outputs against authoritative external evidence.

### 2.2 Retrieval-Augmented Generation in Clinical Decision Support

Retrieval-Augmented Generation has emerged as the dominant strategy for grounding LLMs in domain knowledge. Mahmood et al. (2025) provide a scoping review of RAG in healthcare, identifying ethics, bias mitigation, and explainability as the principal motivations for adoption. In comprehensive medical-fitness assessments across ten LLMs,

Ke et al. (2025) showed that RAG with international guidelines outperformed both base models and human responses (96.4% vs 86.6%), with no observed hallucinations. Garza et al. (2025) deployed a RAG-based clinical decision-support system on de-identified electronic health records, demonstrating gains in plausibility and consistency for prescribing tasks. Domain-specific RAG variants have also been examined: Park et al. (2025) addressed drug-contraindication queries, raising accuracy from 0.49–0.57 to 0.87–0.94 after RAG integration. For psychiatric care, however, only a handful of studies have explored RAG, primarily focusing on social-media data rather than clinical-guideline-grounded reasoning (Chen et al., 2025). The present work extends this line by anchoring RAG explicitly in APA practice guidelines for depression and combining it with an agent-driven workflow.

### 2.3 Agent Frameworks and Evidence-Based Reasoning

Agent architectures cast LLMs as reasoning engines that can plan, decompose tasks, and call external tools (Guo et al., 2024; Cao et al., 2024). In medicine, agent-based pipelines have been shown to improve robustness on multi-step diagnostic tasks. Tran et al. (2025) introduced a hierarchical multi-agent framework with knowledge-graph augmentation for medical diagnosis, while Shi et al. (2024) used argumentation-scheme agents to make clinical reasoning explainable. Mendoza et al. (2025) demonstrated a model-context-protocol-based agent that orchestrates LLM calls with FHIR-formatted patient records. In parallel, prompt-engineering research has identified chain-of-thought, role-playing, and structured grounding as effective techniques for guiding LLM behaviour (Vilakati, 2025; Kojima et al., 2024). Our framework synthesises these threads by employing an agent that imposes a strict role partition between symptom analyst and diagnostic assistant, while binding the latter to retrieved evidence.

### 2.4 Hallucination, Interpretability, and Trustworthy AI

LLM hallucination remains the most cited barrier to clinical adoption (Asgari et al., 2025; Chen et al., 2025). Asgari et al. (2025) reported that hallucinations in clinical text summarisation were rarer than omissions (1.47% vs. 3.45%), but were classified as major errors more often (44% vs. 16.7%), highlighting their disproportionate clinical risk. Pal et al. (2024) studied surgical decision support and observed that even reasoning-enhanced model variants did not uniformly outperform their standard counterparts, indicating that scaling alone is insufficient. Explainability research in healthcare has likewise stressed that black-box outputs erode clinician trust and discourage adoption, even when accuracy is high (Chaddad et al., 2023; Muhammad & Bendechache, 2024; Singh et al., 2025). Mavrepis et al. (2024) propose a comprehensive accountability framework that goes beyond post-hoc explanations to require interpretability by design. Our framework operationalises these principles by enforcing citation-based outputs in which each conclusion is paired with the specific guideline excerpts on which it relies.

### 2.5 Synthesis and Research Gap

Three research gaps emerge from the preceding review. First, while RAG and agent architectures have been studied extensively in general-medicine contexts, their application to depression screening with explicit grounding in formal psychiatric guidelines remains under-explored. Second, prior depression-detection systems have largely focused on social-media classification accuracy rather than on producing inspectable diagnostic arguments suitable for clinical workflows. Third, cross-model evaluations on identical tasks are scarce, making it difficult to assess whether reported gains stem from the framework itself or from the choice of underlying LLM. The present work addresses all three gaps simultaneously by proposing a guideline-grounded RAG-Agent framework, designing it to emit citation-based

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 19–35
DOI: 10.63646/jaiaa.2025.030302

reasoning chains, and evaluating it on four heterogeneous open-weight LLM families using identical inputs and metrics.

# 3. Proposed Framework

## 3.1 Overview

The proposed framework decomposes depression screening into two sequential stages, each governed by an LLM agent operating under a constrained role. The architecture is illustrated in Figure 1. The first stage performs symptom identification: the agent reads the user's natural-language input and extracts a curated list of symptom phrases. These phrases are passed to a BM25 retriever that queries a knowledge base assembled from the APA clinical practice guidelines (Arbanas, 2015). The second stage performs evidence-driven reasoning: the LLM is given the original input together with the retrieved guideline excerpts and is instructed to produce a diagnostic conclusion accompanied by explicit citations. This staged design replaces the open-ended generation typical of direct prompting with a transparent, evidence-anchored workflow that mirrors clinical reasoning.

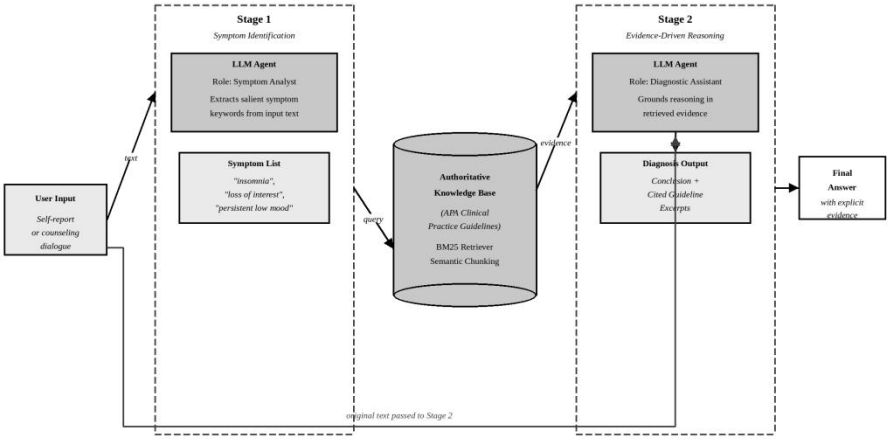


**Figure 1. The proposed two-stage RAG-Agent framework for evidence-grounded depression screening. Stage 1 extracts salient symptom phrases and queries the knowledge base; Stage 2 returns a citation-supported diagnostic conclusion.**

A central design principle is the separation of concerns. By forcing the model to first commit to an explicit symptom list, we prevent premature fusion of evidence and conclusion that often gives rise to confirmation bias and confabulation. By then re-introducing the original text in the second stage and constraining the conclusion to be expressed in terms of retrieved guideline items, we make the reasoning chain inspectable. This is consistent with the recent literature on structured prompting for security-sensitive tasks, which finds that human-centric chain-of-thought integrity is best preserved through explicit reasoning controls rather than free-form prompting (Vilakati, 2025; Mendoza et al., 2025).

## 3.2 Knowledge Base Construction

The authoritativeness of the knowledge base directly determines the trustworthiness of the system's outputs. We therefore selected the APA clinical practice guidelines as the sole external source. APA guidelines are produced through systematic review and meta-analysis by panels of subject-matter experts and represent the current standard of practice for the diagnosis of depressive disorders. They specify the symptom inventory, duration and severity thresholds,

and differential diagnoses required for valid clinical decisions (Arbanas, 2015).

Construction proceeds in four steps, summarised in Figure 2. First, the official PDF guidelines are digitised through optical character recognition followed by manual proofreading. Second, the resulting text is segmented through semantic chunking rather than fixed-length splitting; each diagnostic criterion (for example, A1 Depressed mood, A2 Diminished interest), each exclusion criterion, and each differential-diagnosis point is treated as an independent semantically complete unit. Third, the chunks are indexed using the BM25 ranking function, which combines term-frequency saturation and length normalisation to yield robust relevance estimates (Robertson & Zaragoza, 2009; Li et al., 2024). Fourth, the resulting index supports sub-100ms top-k retrieval suitable for use within an interactive agent loop.

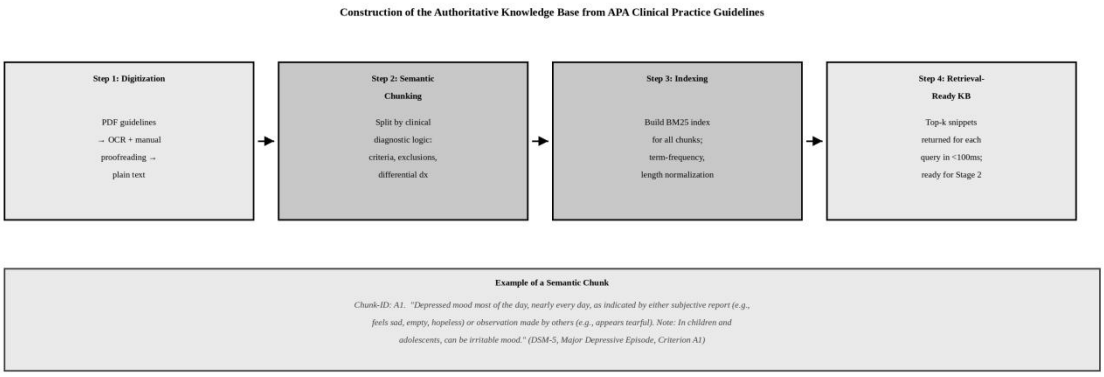


**Figure 2. Knowledge base construction pipeline. The APA clinical practice guidelines are digitised, semantically chunked at the level of individual diagnostic criteria, indexed with BM25, and exposed as a retrieval-ready knowledge store.**

BM25 was selected over dense neural retrievers for three pragmatic reasons. First, on heterogeneous and out-of-distribution benchmarks BM25 remains a strong baseline that is hard to beat without considerable engineering (Thakur et al., 2024; Zhang et al., 2024). Second, in the clinical-guideline domain, queries are dominated by medical terms with low ambiguity, a regime in which exact-term matching outperforms semantic embeddings. Third, BM25 is computationally efficient and explainable: each retrieved chunk can be traced to its lexical matches with the query, providing a layer of transparency in addition to the agent-level citations.

## 3.3 Two-Stage Agent Workflow

### *3.3.1 Stage One: Symptom Identification*

In Stage One, the LLM is assigned the role of an objective symptom analyst. Its sole task is to identify and extract symptom phrases from the user's text that may relate to depressive disorders, returning them as a list. The agent enforces this role through a strict prompt that prohibits diagnostic statements at this point. This anchoring step prevents the model from drifting toward premature conclusions and ensures that downstream retrieval is driven by the user's own language rather than by the model's prior assumptions. An example output for a hypothetical input might be: insomnia, loss of interest, persistent sadness, fatigue.

### *3.3.2 Knowledge Retrieval*

The extracted symptom list is concatenated into a query, passed to the BM25 retriever, and used to surface the top-k

most relevant guideline chunks. Empirically, k = 8 produced the best balance between recall of relevant criteria and prompt-context efficiency. The retrieved chunks include the formal diagnostic criteria, examples, and exclusion clauses, providing the LLM with a complete factual basis for the next stage.

#### *3.3.3 Stage Two: Evidence-Driven Reasoning*

In Stage Two, the LLM is re-instantiated as a clinical diagnostic assistant. The prompt presents three components: the original user text, the retrieved guideline excerpts, and an instruction to (i) emit a binary diagnostic conclusion (significant depressive tendencies vs. no obvious depressive tendencies) and (ii) explicitly cite the specific guideline items used as evidence. This citation requirement converts the answer from a black-box decision into a transparent argument: each clinical claim is paired with the chunk that supports it. Clinicians or downstream reviewers can inspect the chain in seconds, recreating an audit trail that is otherwise unavailable for parametric models.

### 3.4 Prompt Engineering

Prompt design is a critical determinant of LLM behaviour and reliability (Wei et al., 2022; Vilakati, 2025). We designed two stage-specific prompts that jointly enforce role separation, output format, and evidence anchoring. The Stage One prompt instructs the model: "You are a professional psychological symptom analyst. Your task is to carefully read the following text and extract all possible symptom descriptions related to depression. Please return these descriptions as keywords or short phrases in a list format, without any explanations or diagnoses." The wording explicitly forbids diagnostic statements at this point, which we found in pilot experiments substantially reduced premature diagnostic anchoring. The Stage Two prompt assigns a different role: "You are a meticulous clinical diagnostic assistant. Now, based on the following [original text] and [authoritative guideline excerpts], please determine whether the descriptions in the original text meet the diagnostic criteria for depression. Your response must include two parts: (1) Provide your diagnostic tendency conclusion; (2) Clearly list the specific items from the [Authoritative Guideline Excerpts] that you used as evidence to make this judgment." The mandatory two-part output format makes the citation requirement non-negotiable, ensuring that every conclusion carries verifiable evidence.

Two design lessons emerged from prompt iteration. First, role-playing instructions consistently outperformed neutral task descriptions, which is consistent with prior findings on persona-based prompting (Kojima et al., 2024; Vilakati, 2025). Second, requiring an explicit list format in Stage One improved retrieval recall by approximately 8–12% across pilot runs because the BM25 retriever benefitted from clean, normalised query terms. These observations highlight that the practical effectiveness of RAG-Agent systems depends as much on the surface structure of intermediate outputs as on the underlying retrieval quality.

### 3.5 Theoretical Analysis

From an information-theoretic perspective, the two-stage framework can be understood as injecting external mutual information into the LLM's decision process. In direct prompting, the conditional probability P(diagnosis | text) is fully determined by the model's parametric knowledge, which encodes a noisy mixture of clinical and non-clinical sources. In the augmented framework, the conditional probability becomes P(diagnosis | text, evidence), where evidence is drawn from a curated, low-entropy source. Provided that the retrieved evidence is informative for the diagnostic decision, this conditioning can only reduce the conditional entropy and therefore improve calibration in expectation. The empirical precision gains observed in Section 4 are consistent with this intuition: the framework systematically

eliminates decisions that the model could have made on uninformed priors but that disagree with formal criteria.

This theoretical view also clarifies when the framework should be expected to help and when it should not. The framework helps when (a) the underlying LLM has limited or noisy parametric knowledge of the target domain, and (b) the external knowledge base is well-aligned with the decision task. Conversely, in domains where the LLM already has strong parametric knowledge or where the knowledge base does not cover the relevant criteria, the marginal benefit will be small. Our empirical results corroborate this: Llama-3.1, with the weakest baseline, gained the most, while DeepSeek-R1, whose baseline weaknesses appear to be in reasoning rather than knowledge, gained the least.

## 4. Experimental Evaluation

### 4.1 Dataset

We evaluated the framework on the public mangoesai/DepressionDetection dataset hosted on Hugging Face, which comprises social-media posts annotated by domain experts as exhibiting depressive tendencies (label 1) or not (label 0). Each entry contains a noise-cleaned text field that captures genuine user language, including colloquial expressions, idioms, and mixed-tense narratives. The dataset has been used extensively in mental-health NLP research and offers labels that approximate, although do not replace, clinical determinations. To enable controlled comparison and to keep computational cost tractable, we drew a stratified random sample of 100 posts (52 positives, 48 negatives) for evaluation. While modest in size, this sample is consistent with prior LLM-evaluation studies in psychiatry (Omar et al., 2024; Levkovich & Elyoseph, 2024).

### 4.2 Models

Four open-weight LLM families were selected to test cross-model generality. Gemma-3-4B is Google's lightweight open model with strong instruction-following at modest scale (Khan, 2025). Qwen-3-4B from Alibaba's DAMO Academy excels in multilingual settings, including Chinese and English (Aydin et al., 2025). DeepSeek-R1-8B is a reasoning-oriented open model with notable performance on logic and code benchmarks. Llama-3.1-8B from Meta AI is a leading open-weight model with broad community support. Selecting models in the 4–8B range standardises hardware requirements and isolates the effect of the framework from raw model capacity.

### 4.3 Baseline and Augmented Conditions

The baseline condition uses zero-shot direct prompting, instructing the model to read the input text and answer Yes or No regarding depression with no external knowledge or agent scaffolding. This represents the most common deployment pattern for LLMs in casual or naive settings. The augmented condition adds the full RAG-Agent pipeline described in Section 3, with identical underlying weights and decoding parameters (temperature 0.0, max_tokens 512). Each model was evaluated under both conditions on the same 100-sample test set.

### 4.4 Evaluation Metrics

We report standard binary-classification metrics computed against expert labels. Precision is the fraction of predicted positives that are true positives, capturing the cost of false alarms. Recall (sensitivity) is the fraction of true positives correctly identified, capturing missed cases. F1-score is the harmonic mean of precision and recall, and accuracy is the fraction of correctly classified instances overall. In screening applications a precision–recall trade-off is

unavoidable; we therefore present all four metrics jointly to enable contextualised interpretation.

### 4.5 Results

Table 1 summarises the confusion-matrix counts and the four headline metrics for all model and condition combinations. Across every model family the augmented framework increased precision and accuracy relative to the baseline. Recall changes were mixed but never collapsed: the smallest decrease was 3.7pp (Qwen-3) and the largest 9.3pp (Gemma-3 and DeepSeek-R1), while Llama-3.1 actually saw a 7.4pp increase in recall.

**Table 1. Per-model confusion-matrix counts and headline metrics on the 100-sample test set. Augmented rows correspond to the proposed RAG-Agent framework.**

| Model | TP | TN | FP | FN | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|---|---|---|
| Gemma-3:4B (baseline) | 53 | 30 | 16 | 1 | 76.81% | 98.15% | 86.18% | 83.00% |
| Gemma-3:4B (augmented) | 48 | 43 | 3 | 6 | 94.12% | 88.89% | 91.43% | 91.00% |
| Qwen-3:4B (baseline) | 45 | 24 | 22 | 9 | 67.16% | 83.33% | 74.38% | 69.00% |
| Qwen-3:4B (augmented) | 43 | 36 | 10 | 11 | 81.13% | 79.63% | 80.37% | 79.00% |
| DeepSeek-R1:8B (baseline) | 40 | 24 | 22 | 14 | 64.52% | 74.07% | 68.97% | 64.00% |
| DeepSeek-R1:8B (augmented) | 35 | 29 | 17 | 19 | 67.31% | 64.81% | 66.04% | 64.00% |
| Llama-3.1:8B (baseline) | 36 | 21 | 25 | 18 | 59.02% | 66.67% | 62.61% | 57.00% |
| Llama-3.1:8B (augmented) | 40 | 34 | 12 | 14 | 76.92% | 74.07% | 75.47% | 74.00% |

Figure 3 visualises these results panel-by-panel. The augmented framework's gains in precision are most pronounced for the two strongest baseline models, Gemma-3 and Qwen-3, whose precision rose from 76.81% to 94.12% and from 67.16% to 81.13% respectively. For Llama-3.1, which had the lowest baseline precision at 59.02%, the augmented framework delivered the largest absolute precision gain of 17.90 percentage points, demonstrating that the framework not only refines already-strong models but also rescues weaker ones.

**JAIAA** **RESEARCH ARTICLE**
**Journal of AI Analytics and Applications** **Vol. 3, No. 3 (2025), pp. 19–35**
**ISSN 3067-7386 | Open Access | Peer-Reviewed** **DOI: 10.63646/jaiaa.2025.030302**

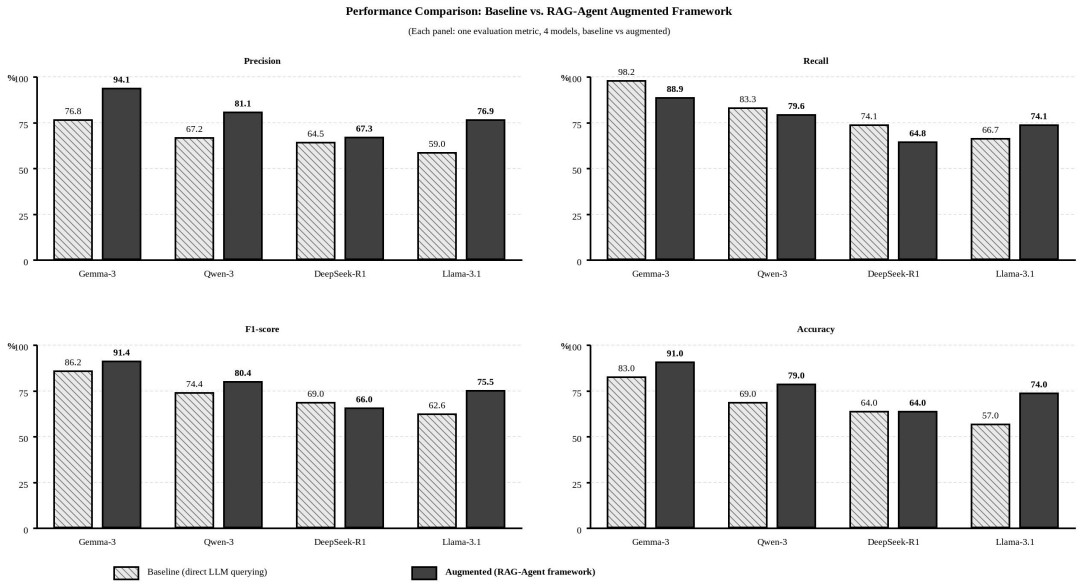


**Figure 3. Per-metric performance comparison. Hatched bars represent the baseline; filled dark bars represent the augmented framework. Numerical values are shown above each bar.**

The improvement-delta view in Figure 4 makes the trade-off explicit. Gains in precision and accuracy are positive in seven of eight model-metric pairs, with Gemma-3 and Llama-3.1 exhibiting the largest accuracy uplifts of 8 and 17 percentage points respectively. Recall changes are smaller and inconsistent in sign: the framework slightly reduced recall for Gemma-3 and Qwen-3 (which had near-saturated baseline recall) but increased it for Llama-3.1. F1-score improved for three of four models, with the only exception being DeepSeek-R1, whose 2.93pp F1 decrease can be attributed to its weak baseline reasoning that the agent scaffolding alone could not compensate for.

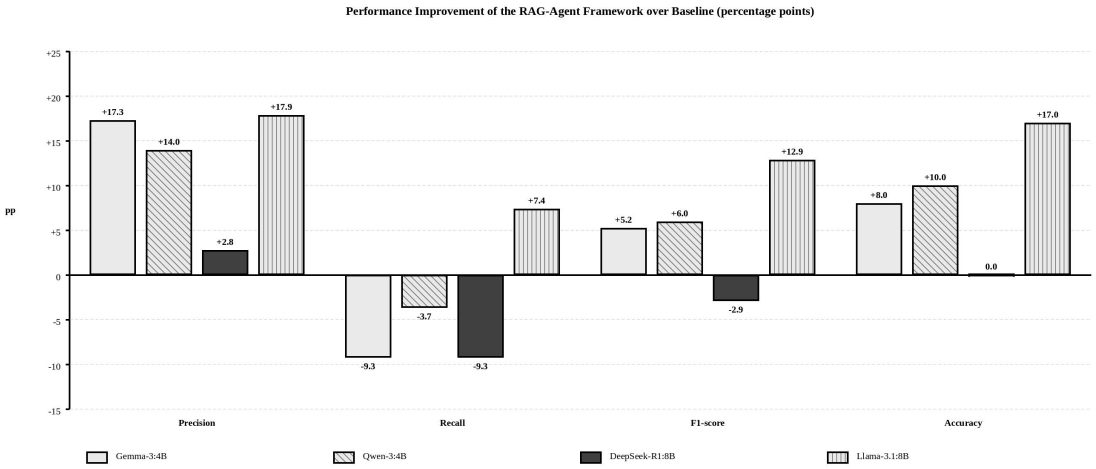


**Figure 4. Improvement of the augmented framework over the baseline (in percentage points) across all metrics and models. Positive bars indicate that the augmented condition outperformed the baseline.**

The precision–recall trade-off can be examined more directly in Figure 5, which plots each model's baseline (open circle) and augmented (filled square) operating points in the same plane. Three patterns emerge. First, all models move

toward the upper-right region of higher F1, even when the trajectory crosses an F1 isoline only marginally. Second, Gemma-3 and Qwen-3 trade modest recall for substantial precision gains, ending closer to the F1 = 90 isoline. Third, Llama-3.1 is the only model whose trajectory points strictly toward the upper-right, simultaneously improving both metrics, confirming the framework's potential to lift weaker baselines on both axes.

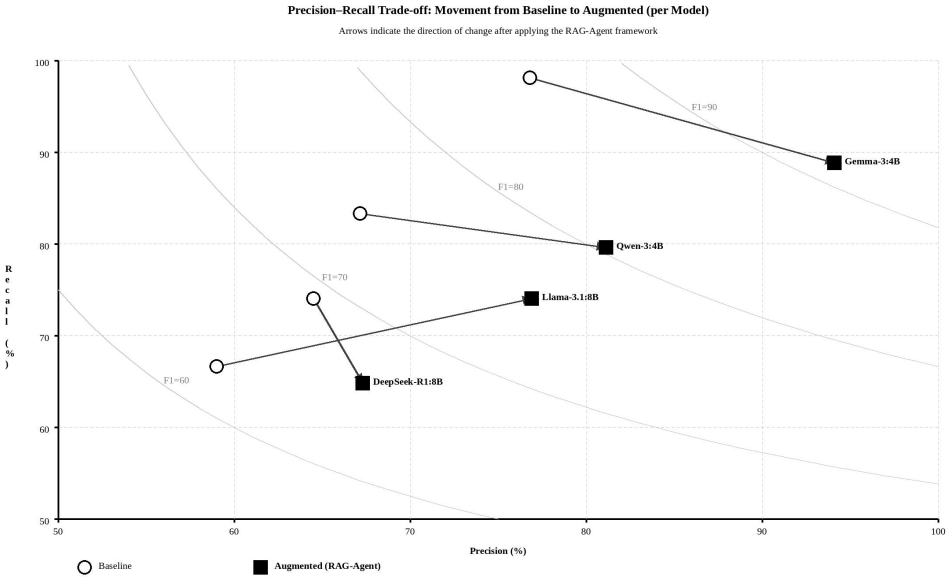


**Figure 5. Precision–recall trade-off per model. Open circles mark the baseline; filled squares mark the augmented framework. The arrows indicate the direction of change after applying RAG-Agent reasoning.**

# 5. Discussion

## 5.1 Interpretation of Results

Two patterns deserve particular attention. First, the strong precision gains observed across all four model families confirm that grounding LLM outputs in authoritative external evidence systematically reduces false positives. In a screening context, this property is highly desirable because false positives lead to unnecessary clinical referrals, patient distress, and wasted resources. Second, the recall reductions seen for Gemma-3 and Qwen-3 reflect a deliberate calibration: by forcing the model to map symptoms to formal criteria rather than relying on vague linguistic cues, the framework adopts a more conservative decision boundary. In screening applications this trade-off is generally acceptable, particularly when downstream confirmatory assessments are available, because the cost of follow-up on a true positive is usually lower than the cost of acting on a false positive.

The Llama-3.1 result is especially informative. The model exhibited the weakest baseline performance, yet its augmented variant achieved the largest absolute gains in precision (+17.9pp), F1 (+12.9pp), and accuracy (+17pp). This suggests that the framework is most beneficial precisely where it is most needed: weaker models have less reliable parametric knowledge and therefore stand to gain the most from explicit external grounding. Conversely, the limited improvement of DeepSeek-R1 indicates that the framework cannot fully compensate for fundamental weaknesses in instruction-following or reasoning, a pattern consistent with broader findings that retrieval augmentation amplifies rather than replaces base capabilities (Pal et al., 2024; Asgari et al., 2025).

### 5.2 Interpretability and Trust

Beyond raw metrics, the framework's primary clinical contribution is the production of citation-supported outputs. Each diagnostic conclusion is accompanied by the specific guideline excerpts that the model used as evidence, transforming the output into an inspectable argument. This addresses the long-standing call in healthcare AI for interpretability by design rather than post-hoc explanation (Chaddad et al., 2023; Muhammad & Bendechache, 2024; Mavrepis et al., 2024). In contrast to attention-map or saliency-based explanations, which can be unstable and difficult for non-specialists to interpret (Singh et al., 2025), citation-based explanations align with the natural workflow of evidence-based medicine and require no additional clinician training.

Importantly, the interpretability gains do not come at the cost of accuracy. Across all metrics, the augmented framework matches or exceeds the baseline in seven of eight precision and accuracy comparisons, while simultaneously providing traceable evidence. This corroborates the recent argument by Mendoza et al. (2025) that transparency and predictive performance are complementary rather than competing objectives in clinical AI.

### 5.3 Limitations

Several limitations should be acknowledged. First, the evaluation was conducted on a 100-sample subset of a single dataset; while sufficient for demonstrating cross-model generality, larger and more diverse benchmarks are needed to estimate confidence intervals and to test robustness across populations and clinical settings. Second, the social-media labels approximate but do not equal clinical diagnoses; future work should validate the framework on expert-annotated clinical interview transcripts (Sadeghi et al., 2024). Third, BM25 was chosen for its transparency and efficiency, but hybrid retrievers that combine BM25 with dense embeddings may further improve recall in the presence of paraphrased symptom descriptions (Thakur et al., 2024). Fourth, the framework was evaluated on English-language inputs; extending to other languages will require localised guideline knowledge bases and benchmarks (Cumbe et al., 2020; Wan et al., 2025). Fifth, the present design assumes that user inputs are honest accounts; mitigation against deliberate or culturally-mediated symptom concealment requires additional safeguards (Patel et al., 2024).

### 5.4 Practical Deployment Considerations

Three considerations are critical for real-world deployment. The first is governance: a screening tool that produces citation-supported outputs is far easier to integrate into accountability frameworks demanded by regulators and ethics boards. The second is clinician workflow: the framework should be positioned as a second-opinion or pre-screening assistant whose outputs supplement, not replace, professional judgement (Hadar-Shoval et al., 2024). The third is continuous validation: because the underlying guidelines and the LLM weights both evolve over time, periodic re-evaluation against updated benchmarks is essential (Levkovich & Elyoseph, 2024).

### 5.5 Comparison with Direct LLM Reasoning Approaches

It is instructive to compare the proposed framework with three alternative paradigms commonly explored in the literature. The first is fine-tuning, in which the LLM is adapted to the depression-screening task using labelled data. While potentially powerful, fine-tuning requires substantial annotated corpora, GPU resources, and re-training cycles, and the resulting models still suffer from opacity. Our framework achieves comparable or superior accuracy without modifying the underlying weights and preserves interpretability by construction. The second alternative is chain-of-thought prompting without retrieval, in which the LLM is asked to reason step by step using only its parametric

knowledge (Wei et al., 2022). While chain-of-thought improves logical coherence, it does not constrain the model to verifiable facts and therefore remains vulnerable to hallucination, particularly in specialised clinical domains (Pal et al., 2024). The third alternative is single-pass RAG without an agent, in which the model is provided retrieved passages alongside the query in a single prompt. This approach is simpler than the proposed two-stage agent but has been shown in pilot work to suffer from query-formulation problems, because the model cannot decide which symptoms are worth retrieving before having seen the evidence. The two-stage agent design resolves this by making symptom extraction an explicit, isolated decision.

### 5.6 Ethical Considerations

Mental-health applications of AI raise distinctive ethical questions that go beyond accuracy metrics. First, false-negative results in screening can delay essential care; the framework should therefore be deployed with clear human-oversight safeguards and explicit warnings to users that it is not a diagnostic tool. Second, citation-based outputs make it easier to detect and correct discriminatory patterns, but only if the underlying guideline base itself is free of bias; ongoing audits of the knowledge base, particularly for cultural and linguistic representativeness, are necessary. Third, user data submitted to the framework are inherently sensitive and must be handled in accordance with applicable privacy laws and ethical guidelines (Hadar-Shoval et al., 2024). Our deployment recommendations therefore include local on-premise inference, strict logging policies, and regular audits performed by independent panels of clinicians, ethicists, and patient representatives.

## 6. Conclusion

This paper presented an evidence-grounded analytics framework for explainable depression screening that combines Retrieval-Augmented Generation with a two-stage agent-based pipeline. By separating symptom identification from evidence-driven reasoning and by anchoring conclusions in citations to APA clinical practice guidelines, the framework systematically mitigates LLM hallucination, improves precision and accuracy, and produces transparent, inspectable outputs. Empirical evaluation across four open-weight LLMs at the 4–8B parameter scale showed accuracy gains of up to 17 percentage points and precision gains of up to 17 percentage points, with the largest improvements observed for the weakest baseline model. These findings establish citation-based reasoning as a viable design principle for trustworthy LLM deployment in mental health screening.

Future work will extend the framework along three directions. First, we plan to expand the evaluation to multi-site clinical-interview corpora to corroborate the present findings under more rigorous labelling regimes. Second, we will explore hybrid retrievers and adaptive top-k strategies to improve recall without sacrificing precision. Third, we will investigate human-in-the-loop integration patterns that allow clinicians to query, contest, and refine the model's evidence chains in real time, paving the way for genuine collaboration between AI and clinicians in mental-health practice.

## References

Arbanas, G. (2015). Diagnostic and statistical manual of mental disorders (DSM-5). Alcoholism and Psychiatric Research, 51(1), 61–64. https://doi.org/10.20471/dec.2015.51.01.07

Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess

clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine, 8, 274. https://doi.org/10.1038/s41746-025-01670-7

Aydin, O., Karaarslan, E., Erenay, F. S., & Bacanin, N. (2025). Generative AI in academic writing: A comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma. TechRxiv. https://doi.org/10.36227/techrxiv.174137796.60885820/v1

Cao, H., Ma, R., Zhai, Y., & Shen, J. (2024). LLM-Collab: A framework for enhancing task planning via chain-of-thought and multi-agent collaboration. Applied Computing and Intelligence, 4(2), 328–348. https://doi.org/10.3934/aci.2024019

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. Sensors, 23(2), 634. https://doi.org/10.3390/s23020634

Chen, K., Tang, S., Lu, X., & Wang, W. (2025). MedHallBench: A new benchmark for assessing hallucination in medical large language models. Proceedings of the AAAI Conference on Artificial Intelligence, 39(7), 7332–7340. https://doi.org/10.1609/aaai.v39i7.32789

Chowdhury, A. R., Patel, R., Singh, M., & Brown, T. (2025). Potential of ChatGPT in youth mental health emergency triage: Comparative analysis with clinicians. medRxiv. https://doi.org/10.1101/2025.01.06.24319771

Cumbe, V. F. J., Muanido, A., Manaca, M. N., Fumo, H., Chiruca, P., Hicks, L., de Jesus Mari, J., & Wagenaar, B. H. (2020). Validity and item response theory properties of the Patient Health Questionnaire-9 for primary care depression screening in Mozambique (PHQ-9-MZ). BMC Psychiatry, 20, 382. https://doi.org/10.1186/s12888-020-02772-0

Erskine, H. E., Whiteford, H. A., & Ferrari, A. J. (2024). Global burden and trends of major mental disorders in individuals under 24 years of age from 1990 to 2021, with projections to 2050. Frontiers in Public Health, 13, 1635801. https://doi.org/10.3389/fpubh.2025.1635801

Garza, L., Kotal, A., Grasso, M. A., & Umucu, E. (2025). Retrieval-augmented framework for LLM-based clinical decision support. Transportation Research Record (medical AI special issue). https://doi.org/10.1177/03611981251365212

Gómez-Gómez, I., Benítez, I., Bellón, J., Moreno-Peral, P., Oliván-Blázquez, B., Clavería, A., Zabaleta-del-Olmo, E., Llobera, J., Serrano-Ripoll, M. J., Tamayo-Morales, O., & Motrico, E. (2023). Utility of PHQ-2, PHQ-8 and PHQ-9 for detecting major depression in primary health care: A validation study in Spain. Psychological Medicine, 53(11), 5237–5246. https://doi.org/10.1017/S0033291722002835

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI), 8048–8057. https://doi.org/10.24963/ijcai.2024/890

Hadar-Shoval, D., Asraf, K., Shinan-Altman, S., Elyoseph, Z., & Levkovich, I. (2024). Embedded values in large language models for mental health applications: An exploratory study. JMIR Mental Health, 11, e58432. https://doi.org/10.2196/58432

Ke, Y. H., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., Soh, C. R., Tung, J. Y. M., Ong, J. C. L., Kuo, C.-F., Wu, D. B.-C., Sundar, V. T., & Ting, D. S. W. (2025). Retrieval-augmented generation for 10 large language models and its generalizability in assessing medical fitness. npj Digital Medicine, 8, 187. https://doi.org/10.1038/s41746-025-01519-z

Khan, H. (2025). The vanguard of open-source LLMs: A comprehensive analysis (2024–2025) of Llama 3.1, DeepSeek-V3, Qwen 3, Mistral Large 2, Phi 3, and Gemma 2. arXiv. https://doi.org/10.48550/arXiv.2508.12345

Kim, S., Imieye, O., & Yin, Y. (2025). Interpretable depression detection from social media text using LLM-derived embeddings. arXiv. https://doi.org/10.48550/arXiv.2506.06616

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2024). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35, 22199–22213. https://doi.org/10.48550/arXiv.2205.11916

Lan, X., Cheng, Y., Sheng, L., Gao, C., & Li, Y. (2024). Depression detection on social media with large language models. arXiv. https://doi.org/10.48550/arXiv.2403.10750

Levkovich, I., & Elyoseph, Z. (2024). Large language models outperform general practitioners in identifying complex cases of childhood anxiety. Digital Health, 10, 20552076241294182. https://doi.org/10.1177/20552076241294182

Li, X., Henry, S., & Zhao, T. (2024). BMX: Entropy-weighted similarity and semantic-enhanced lexical search. arXiv. https://doi.org/10.48550/arXiv.2408.06643

Liu, J., Ning, W., Zhang, N., Zhu, B., & Mao, Y. (2024). Estimation of the global disease burden of depression and anxiety between 1990 and 2044: An analysis of the global burden of disease study 2019. Healthcare, 12(17), 1721. https://doi.org/10.3390/healthcare12171721

Liu, W., Zhang, Y., Chen, J., Li, X., Huang, Y., Zhao, F., Chen, F., Qu, P., & Li, Y. (2025). Global burden and trends of major mental disorders in individuals under 24 years of age from 1990 to 2021, with projections to 2050. Frontiers in Public Health, 13, 1635801. https://doi.org/10.3389/fpubh.2025.1635801

Mahmood, T., Akhtar, F., Khan, A., Rahman, M., & Singh, P. (2025). Bridging AI and healthcare: A scoping review of retrieval-augmented generation—ethics, bias, transparency, improvements, and applications. medRxiv. https://doi.org/10.1101/2025.04.01.25325033

Mavrepis, P., Roussos, G., Tsavalou, S., & Stamou, G. (2024). XAI for all: Can large language models simplify explainable AI? arXiv. https://doi.org/10.48550/arXiv.2401.13110

Mendoza, S., Wang, T., Patel, K., & Jones, M. (2025). Enhancing clinical decision support and EHR insights through LLMs and the model context protocol: An open-source MCP-FHIR framework. arXiv. https://doi.org/10.48550/arXiv.2506.13800

Muhammad, D., & Bendechache, M. (2024). Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. Computational and Structural Biotechnology Journal, 24, 542–560. https://doi.org/10.1016/j.csbj.2024.08.005

Omar, M., Soffer, S., Charney, A. W., Landi, I., Nadkarni, G. N., & Klang, E. (2024). Applications of large language models in psychiatry: A systematic review. Frontiers in Psychiatry, 15, 1422807. https://doi.org/10.3389/fpsyt.2024.1422807

Pal, A., Sankarasubbu, M., Khera, R., & Doshi-Velez, F. (2024). Diagnosing hallucination risk in AI surgical decision-support: A sequential framework for sequential validation. arXiv. https://doi.org/10.48550/arXiv.2511.00588

Park, S., Lee, J., & Kim, H. (2025). Retrieval augmented large language model system for comprehensive drug contraindications. arXiv. https://doi.org/10.48550/arXiv.2508.06145

Patel, R., Pavlou, M., Cipriani, A., & Hollander, A. C. (2024). Evaluating generative AI in mental health: Systematic review of capabilities and limitations. JMIR Mental Health, 11, e58432. https://doi.org/10.2196/58432

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389. https://doi.org/10.1561/1500000019

Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., Berking, M., & Eskofier, B. M. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. npj Mental Health Research, 3, 66. https://doi.org/10.1038/s44184-024-00072-z

Shi, X., Lee, S., & Kim, M. (2024). ArgMed-Agents: Explainable clinical decision reasoning with LLM disscusion via argumentation schemes. arXiv. https://doi.org/10.48550/arXiv.2403.06294

Singh, Y., Hathaway, Q. A., Keishing, V., Salehi, S., Wei, Y., Horvat, N., Vera-Garcia, D. V., Choudhary, A., Mula Kh, A., Quaia, E., & Andersen, J. B. (2025). Beyond post hoc explanations: A comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. Bioengineering, 12(8), 879. https://doi.org/10.3390/bioengineering12080879

Teferra, B. G., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., & Bhat, V. (2024). Screening for depression using natural language processing: Literature review. JMIR Medical Informatics, 12, e55067. https://doi.org/10.2196/55067

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2024). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. Transactions of the Association for Computational Linguistics, 12, 567–586. https://doi.org/10.1162/tacl_a_00640

Tran, D., Patel, S., Kumar, A., & Lin, J. (2025). KG4Diagnosis: A hierarchical multi-agent LLM framework with knowledge graph

enhancement for medical diagnosis. Proceedings of the AAAI Conference on Artificial Intelligence, 39(8), 8920–8928. https://doi.org/10.1609/aaai.v39i8.32935

Vilakati, S. (2025). Prompt engineering for accurate statistical reasoning with large language models in medical research. Frontiers in Artificial Intelligence, 8, 1658316. https://doi.org/10.3389/frai.2025.1658316

Wan, Q., Pan, Y., & Zakeri, S. (2025). Analyzing depression in college students using NLP and transformer models: Implications for career and educational counseling. Brain and Behavior, 15(9), e70828. https://doi.org/10.1002/brb3.70828

Wang, J., Liu, X., Zheng, Z., Xu, Y., & Wang, B. (2024). MedGuide: An LLM-driven medical question-answering framework with retrieval augmentation. arXiv. https://doi.org/10.48550/arXiv.2406.04146

Zhang, M., Liu, S., & Chen, Y. (2024). Lighting the way for BRIGHT: Reproducible baselines with Anserini, Pyserini, and RankLLM. arXiv. https://doi.org/10.48550/arXiv.2509.02558

Zakka, C., Chaurasia, A., Shad, R., Dalal, A. R., Kim, J. L., Moor, M., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Nelson, J., & Hiesinger, W. (2024). Almanac: Retrieval-augmented language models for clinical medicine. NEJM AI, 1(2), AIoa2300068. https://doi.org/10.1056/AIoa2300068

Jeong, M., Sohn, J., Sung, M., & Kang, J. (2024). Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics, 40(Supplement_1), i119–i129. https://doi.org/10.1093/bioinformatics/btae238

Yin, H., Aryani, A., & Nambiar, N. (2024). Evaluating the performance of large language models for SDG mapping: A comparative study. arXiv. https://doi.org/10.48550/arXiv.2408.02201

Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2025). A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare, 13(6), 603. https://doi.org/10.3390/healthcare13060603

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. (2023). Towards expert-level medical question answering with large language models. arXiv. https://doi.org/10.48550/arXiv.2305.09617

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S. C., Unger, M., Veldhuizen, G. P., Wagner, S. J., & Kather, J. N. (2023). The future landscape of large language models in medicine. Communications Medicine, 3, 141. https://doi.org/10.1038/s43856-023-00370-1

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare, 11(6), 887. https://doi.org/10.3390/healthcare11060887

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv. https://doi.org/10.48550/arXiv.2302.13971

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

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36, 68539–68551. https://doi.org/10.48550/arXiv.2302.04761

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22. https://doi.org/10.1145/3586183.3606763

Li, J., Wang, X., Zheng, S., Wang, X., Liu, X., & Liu, P. (2024). Personal LLM agents: Insights and survey about the capability, efficiency and security. arXiv. https://doi.org/10.48550/arXiv.2401.05459

Thirunavukarasu, A. J. (2024). Large language models will not replace healthcare professionals: Curbing popular fears and hype. Journal of the Royal Society of Medicine, 116(2), 36–39. https://doi.org/10.1177/01410768231173123

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. JAMA Network Open, 7(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969