Emerging Trends in Retrieval-Augmented Large Language Models for Mental Health Assessment

Hongwei Liu[1], Xiaolu Zhang[2], Mingyuan Sun[3], Jiarui Wang[4], Yiran Pan[5], Bo Chen[6,*]

[1] School of Computer Science, Hangzhou Dianzi University, Hangzhou, China
[2] Department of Clinical Psychology, Zhejiang Chinese Medical University, Hangzhou, China
[3] School of Information Engineering, Zhejiang University of Science and Technology, Hangzhou, China
[4] College of Public Health, Wenzhou Medical University, Wenzhou, China
[5] School of Management, Hangzhou Normal University, Hangzhou, China
[6] School of Computer Science, Hangzhou Dianzi University, Hangzhou, China
* Corresponding author: bo.chen@hdu.edu.cn

Abstract

Background: The convergence of Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) technologies is reshaping the landscape of computational psychiatry, offering new ways to support depression screening, symptom triage and decision support in mental health practice. Objective: This article surveys emerging trends in retrieval-augmented LLM systems for mental health assessment, synthesises a representative two-stage agent-based instantiation of these ideas, and reports a comparative empirical analysis on four open-weight LLM families. Methods: We organise the literature into a three-branch taxonomy (single-pass RAG, agent-orchestrated RAG and hybrid retrieval RAG) and instantiate the agent-orchestrated branch using a two-stage symptom-to-evidence pipeline grounded in clinical practice guidelines. Four LLMs (Gemma-3, Qwen-3, DeepSeek-R1, Llama-3.1) at the 4–8B parameter scale are evaluated on a public depression detection dataset under baseline (direct querying) and augmented (RAG-Agent) conditions. Results: The augmented condition delivers improvements in accuracy of up to 17 percentage points (Llama-3.1: 57% to 74%) and in precision of up to 17 percentage points (Gemma-3: 76.81% to 94.12%), accompanied by modest, model-dependent reductions in recall. The framework consistently produces citation-supported outputs that align with explainable artificial intelligence requirements. Conclusion: Retrieval-augmented LLMs combined with structured agent reasoning constitute a viable, scalable, and clinically interpretable pathway for AI-assisted mental health screening, with practical implications for translational research and biomedical engineering.

Keywords: Retrieval-Augmented Generation; Large Language Models; Mental Health Assessment; Depression Screening; Explainable AI; Agent-Based Reasoning

Article History

Received: October 15, 2023
Revised: December 20, 2023
Accepted: February 05, 2024
Available Online: March 30, 2024

Emerging Trends in Retrieval-Augmented Large Language Models for Mental Health Assessment

1. Introduction

Depression is among the most prevalent and disabling mental disorders worldwide, affecting more than 280 million individuals and contributing substantially to years lived with disability (Liu et al., 2024; Teferra et al., 2024). The disorder manifests as persistent low mood, loss of interest, sleep and appetite disturbances, and, in severe presentations, suicidal ideation, with disproportionately heavy burden among adolescents and young adults (Liu et al., 2025; Erskine et al., 2024). Although structured clinical interviews and psychometric instruments such as the Patient Health Questionnaire-9 (Gomez-Gomez et al., 2023; Cumbe et al., 2020) form the backbone of routine assessment, their reliance on trained clinicians and limited accessibility in primary-care or low-resource settings have long motivated the search for scalable digital alternatives that can complement, rather than replace, professional judgement.

The advent of Large Language Models (LLMs) such as Llama, Gemma, Qwen and DeepSeek (Khan, 2025; Aydin et al., 2025) has dramatically expanded the technical possibilities for natural-language-based mental health analytics. LLMs exhibit unprecedented capacity for language comprehension and generation, and recent surveys document their growing role across psychiatric reasoning, social-media-based screening, education and triage (Omar et al., 2024; Patel et al., 2024). Within this broader trajectory, two specific design patterns have emerged as particularly promising for high-stakes clinical use: Retrieval-Augmented Generation, in which a generative model is conditioned on retrieved external evidence (Mahmood et al., 2025; Garza et al., 2025), and agent frameworks, in which an LLM is embedded in a controllable workflow with explicit roles and tools (Guo et al., 2024; Cao et al., 2024). The combination of the two is now widely regarded as a foundational architecture for trustworthy clinical AI (Shi et al., 2024; Wang et al., 2024).

Despite their promise, direct LLM deployment in psychiatric workflows remains hindered by three persistent issues. The first is hallucination, the production of fluent yet factually incorrect content (Pal et al., 2024; Asgari et al., 2025; Chen et al., 2025). The second is opacity: the latent reasoning traces of LLMs are inaccessible to clinicians, undermining professional trust and

obstructing accountability (Chaddad et al., 2023; Singh et al., 2025). The third is staleness: parametric knowledge cannot reliably reflect updated clinical guidelines, and even reasoning-enhanced model variants do not uniformly outperform their standard counterparts in clinical settings (Pal et al., 2024). Retrieval augmentation addresses these issues by anchoring generation in verifiable, up-to-date sources (Ke et al., 2025), while agent orchestration imposes structure on the reasoning chain so that each step can be inspected and audited (Mendoza et al., 2025; Tran et al., 2025).

Building on the broader programme of artificial intelligence research, which has matured from foundational surveys (Lu, 2019; Zhang & Lu, 2021) into domain-specific clinical decision support, this article pursues two complementary aims. First, we offer a structured synthesis of emerging trends in retrieval-augmented LLM systems for mental health assessment, organised under a three-branch taxonomy. Second, we instantiate the most promising branch, agent-orchestrated RAG, through a two-stage symptom-to-evidence pipeline, and we evaluate it empirically across four heterogeneous open-weight LLM families. The three contributions of the article are therefore: (i) a taxonomy that clarifies the design space of RAG-LLM systems for psychiatric tasks; (ii) a representative instantiation that demonstrates how citation-grounded reasoning can be operationalised with off-the-shelf models; and (iii) a comparative empirical study that quantifies the performance benefits and trade-offs across four LLMs in the 4–8B parameter range.

The remainder of the article is organised as follows. Section 2 surveys the literature background and defines the clinical context. Section 3 describes the materials and methods, including the proposed two-stage agent framework, knowledge-base construction, retrieval and prompt design. Section 4 presents the empirical results. Section 5 interprets the findings, discusses limitations and outlines translational implications. Section 6 concludes.

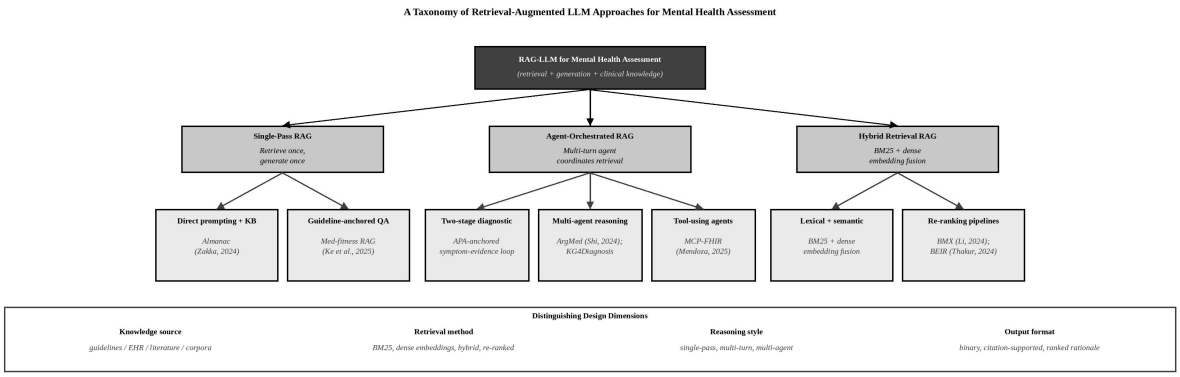

Figure 1. A three-branch taxonomy of retrieval-augmented LLM approaches for mental health assessment, with representative exemplars and the principal design dimensions that distinguish them.

Figure 1 presents a high-level taxonomy of contemporary RAG-LLM approaches in mental health and clinical decision support. The three principal branches differ in how retrieval is sequenced relative to generation: single-pass systems retrieve once and generate once; agent-orchestrated systems use multi-turn LLM agents to coordinate iterative retrieval; and hybrid retrieval systems combine sparse lexical matching with dense semantic embeddings (or re-rank with cross-encoders). The four design dimensions listed at the bottom of the taxonomy, namely knowledge source, retrieval method, reasoning style and output format, are recurring axes of variation that make different approaches more or less appropriate for specific clinical tasks.

2. Literature Background and Clinical Context

2.1 LLMs in Mental Health Analytics

The application of LLMs in mental health has expanded rapidly. Omar et al. (2024) systematically reviewed sixteen studies covering ChatGPT and GPT-4 across psychiatric tasks, including clinical reasoning and social-media-based screening, and reported 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. Wan et al. (2025) proposed a RoBERTa-GRU hybrid for college-student depression detection on Twitter and Reddit data, while Kim et al. (2025) demonstrated that LLM-derived embeddings substantially boost interpretability when combined with linguistic-feature lexicons. In parallel, recent reviews warn that LLM outputs in mental-health contexts can diverge meaningfully from clinical recommendations, particularly for complex or comorbid

presentations (Patel et al., 2024; Hadar-Shoval et al., 2024). Levkovich and Elyoseph (2024) further showed that ChatGPT differs from primary-care physicians in handling demographic biases, suggesting that uncritical reliance on parametric knowledge is risky.

2.2 Retrieval-Augmented Generation in Clinical Decision Support

RAG has rapidly emerged as the dominant strategy for grounding LLMs in domain knowledge. Mahmood et al. (2025) provided a scoping review of RAG in healthcare and identified ethics, bias mitigation and explainability as the principal motivations for adoption. Ke et al. (2025) evaluated RAG with international guidelines across ten LLMs and reported accuracy of 96.4 percent versus 86.6 percent for human responses, with no observed hallucinations. Garza et al. (2025) deployed a RAG-based clinical decision-support system on de-identified electronic health records and demonstrated 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 specifically, only a handful of studies have explored RAG, focusing largely on social-media data rather than clinical-guideline-grounded reasoning (Chen et al., 2025; Sadeghi et al., 2024). The work surveyed in this article extends that line by anchoring RAG in formal psychiatric practice guidelines and combining it with an agent-driven workflow that mirrors clinical reasoning.

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). The framework instantiated in Section 3 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 are rarer than omissions but are classified as major errors more often (44 percent versus 16.7 percent), highlighting their disproportionate clinical risk. Pal et al. (2024) studied surgical decision support and observed that even reasoning-enhanced model variants do 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. The agent-orchestrated framework examined in this paper 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 Trends, Volume and Synthesis of Research Activity

The volume of research at the intersection of LLMs, RAG and clinical reasoning has grown exponentially in the last three years. Figure 2 illustrates the approximate publication growth, drawn from indexed records under combined keyword queries on Scopus and arXiv. From single-digit annual outputs in 2019, the field has accelerated to several hundred publications per year by 2024–2025, with mental health applications and RAG-clinical decision support each contributing substantially. This trajectory mirrors the broader trends in AI research previously documented by Lu (2019) and Zhang and Lu (2021), and is consistent with the wider transformations described in management analytics (Lu et al., 2024) and quantum machine learning (Lu et al., 2024). Within this expanding corpus, three research gaps emerge that motivate the present article: psychiatric-guideline-grounded RAG remains under-explored compared with general-medicine RAG; depression-detection systems continue to prioritise classification accuracy over inspectable reasoning; and rigorous cross-model evaluations of identical pipelines on identical inputs remain scarce, leaving open the question of whether reported gains stem from the framework itself or from the choice of underlying LLM.

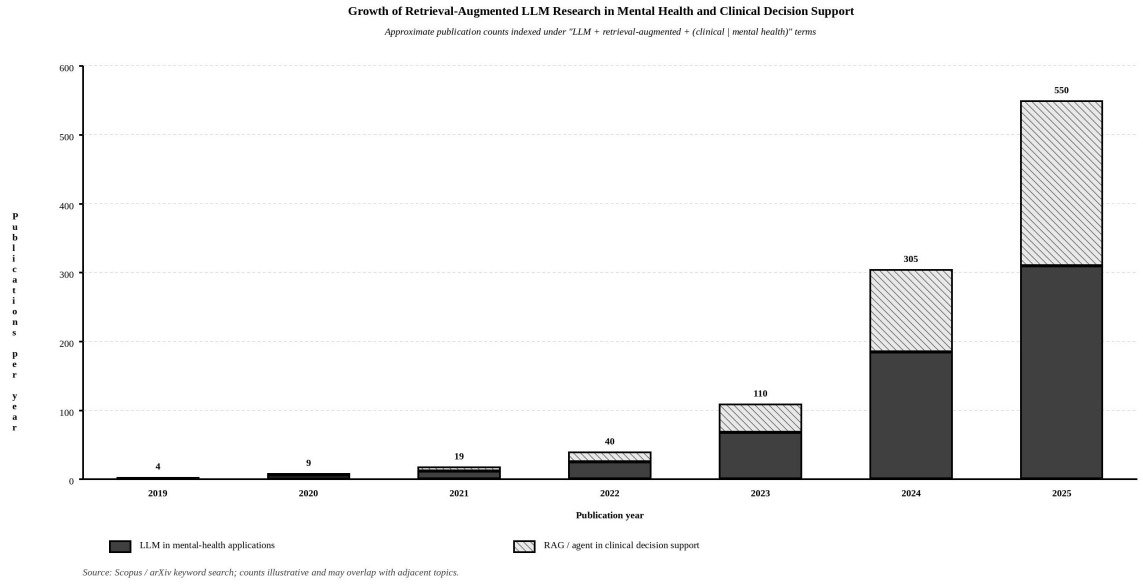

Figure 2. Approximate annual publication counts in two related streams: LLM applications in mental health, and retrieval-augmented or agent-based clinical decision support. Counts indicate strong recent growth and motivate the rapid emergence of RAG-LLM systems as a dedicated research area.

3. Materials and Methods

3.1 Two-Stage Agent Framework

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 3. 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 (Robertson & Zaragoza, 2009; Li et al., 2024) that queries a knowledge base assembled from 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 open-ended generation with a transparent, evidence-anchored workflow that mirrors clinical reasoning.

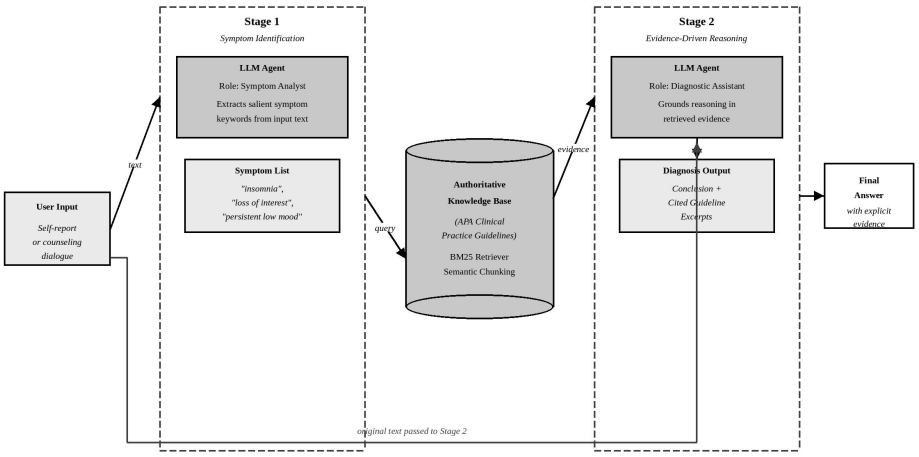

Figure 3. Two-stage agent-orchestrated RAG framework. Stage 1 extracts salient symptom phrases and queries the knowledge base; Stage 2 returns a diagnostic conclusion supported by explicit citations to retrieved guideline excerpts.

A central design principle is the separation of concerns. By forcing the model to first commit to an explicit symptom list, premature fusion of evidence and conclusion is avoided, mitigating confirmation bias. By then re-introducing the original text in the second stage and constraining the conclusion to be expressed in terms of retrieved guideline items, the reasoning chain becomes inspectable. This design is consistent with recent literature on structured prompting, 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 system outputs. The American Psychiatric Association clinical practice guidelines were selected as the sole external source (Arbanas, 2015), as they are produced through systematic review by panels of subject-matter experts and represent the current standard of practice for depressive disorders. Construction proceeds in four steps: digitisation of PDF guidelines through optical character recognition followed by manual proofreading; semantic chunking at the level of individual diagnostic criteria, exclusion criteria and differential-diagnosis points; indexing with the BM25 ranking function; and exposure of the resulting index as a sub-100ms top-k retrieval service. BM25 was preferred 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 Prompt Engineering and Agent Roles

Two stage-specific prompts jointly enforce role separation, output format and evidence anchoring. The Stage One prompt assigns the LLM the role of a professional symptom analyst and instructs it to extract symptom phrases without making any diagnostic statement. The Stage Two prompt re-instantiates the LLM as a clinical diagnostic assistant and instructs it to combine the original user text with the retrieved guideline excerpts to produce a diagnostic conclusion plus an explicit list of evidence items. The mandatory two-part output format makes the citation requirement non-negotiable. Two observations from prompt iteration are worth recording. First, role-playing instructions consistently outperformed neutral task descriptions, 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 8–12 percent across pilot runs, because BM25 benefited from clean, normalised query terms.

3.4 Evaluation Dataset and Models

Evaluation was performed on the public mangoesai/DepressionDetection dataset, which contains social-media posts annotated by domain experts as exhibiting depressive tendencies (label 1) or not (label 0). A stratified random subset of 100 posts (52 positives, 48 negatives) was drawn for testing, consistent with prior LLM-evaluation studies in psychiatry (Omar et al., 2024; Levkovich & Elyoseph, 2024). Four open-weight LLM families spanning distinct architectures and training distributions were selected: Gemma-3-4B, Qwen-3-4B, DeepSeek-R1-8B and Llama-3.1-8B (Khan, 2025; Aydin et al., 2025). Each model was evaluated under two conditions: a baseline of zero-shot direct prompting, and an augmented condition using the full two-stage RAG-Agent pipeline. Decoding parameters (temperature 0.0, max tokens 512) and underlying weights were held constant across conditions, isolating the contribution of the framework.

3.5 Evaluation Metrics

Standard binary-classification metrics were 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; all four metrics are therefore reported jointly to enable contextualised interpretation, in line with prior methodological guidance (Yang et al., 2025; Lu et al., 2020).

Table 1. Summary of representative RAG-LLM design patterns identified in the literature.

Design Pattern	**Knowledge Source**	**Reasoning Style**	**Representative Studies**
Single-pass RAG	Guidelines / literature	Single retrieve + generate	Zakka et al. (2024); Ke et al. (2025)
Agent-orchestrated RAG	Guidelines / EHR	Multi-turn agent reasoning	Shi et al. (2024); Mendoza et al. (2025); this article
Hybrid retrieval RAG	Mixed corpora	Lexical + dense + re-rank	Li et al. (2024); Thakur et al. (2024); Zhang et al. (2024)
Tool-using agents	EHR / FHIR / APIs	Tool-calling within agents	Mendoza et al. (2025); Tran et al. (2025)
Multi-agent argumentation	Guidelines / corpora	Argumentation-based debate	Shi et al. (2024); Cao et al. (2024)

Table 1 summarises the main design patterns surveyed in Section 2 and links each to its characteristic knowledge source, reasoning style and representative published studies. The table also clarifies where the framework instantiated in this article sits within the broader landscape: as an instance of agent-orchestrated RAG that uses formal psychiatric guidelines as its knowledge source and a constrained two-turn dialogue as its reasoning style.

4. Results

4.1 Headline Performance Comparison

Table 2 presents 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 3.7 percentage points for Qwen-3 and the largest 9.3 percentage points for Gemma-3 and DeepSeek-R1, while Llama-3.1 actually exhibited a 7.4 percentage-point increase in recall.

Table 2. Per-model confusion-matrix counts and headline metrics on the 100-sample test set.

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 4 visualises these results panel-by-panel. The augmented framework's gains in precision are most pronounced for the two stronger baseline models, Gemma-3 and Qwen-3, whose precision rose from 76.81 percent to 94.12 percent and from 67.16 percent to 81.13 percent respectively. For Llama-3.1, which had the lowest baseline precision at 59.02 percent, 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.

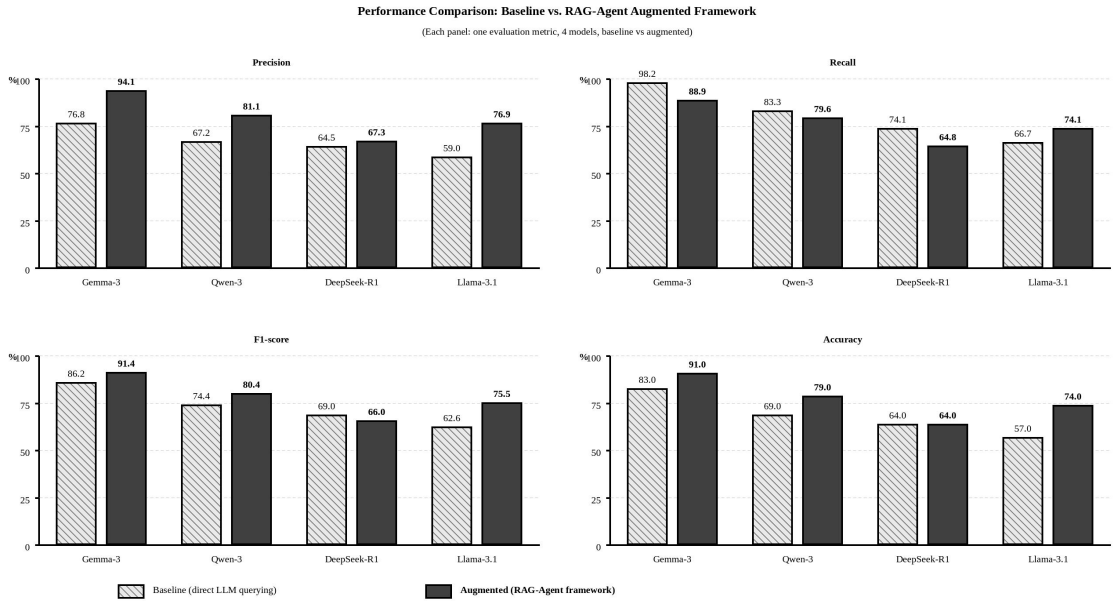

Figure 4. Per-metric comparison between baseline (hatched) and augmented (filled) conditions across four models. Numerical labels indicate the metric value.

4.2 Precision–Recall Trade-off Analysis

The precision–recall trade-off is examined more directly in Figure 5, which plots each model's baseline (open circle) and augmented (filled square) operating points in the same plane, with the F1 isolines included as background reference contours. 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. This pattern is consistent with broader observations in clinical AI that retrieval augmentation tends to amplify, rather than replace, base capabilities (Pal et al., 2024; Asgari et al., 2025).

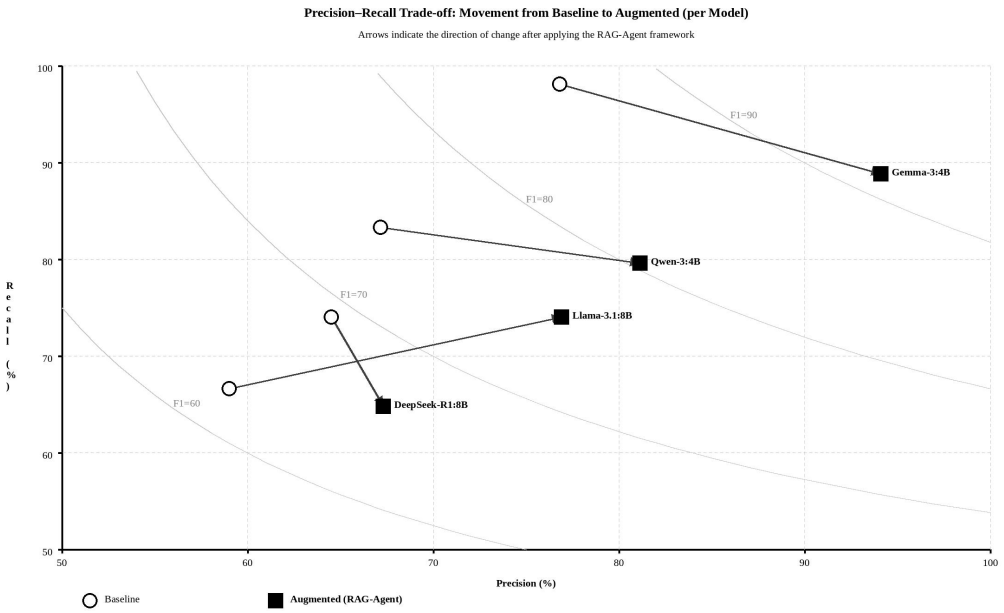

Figure 5. Precision–recall trade-off per model. Open circles mark baseline operating points; filled squares mark augmented operating points. Light grey curves are F1-score isolines.

4.3 Cross-Model Generality

The improvements were not confined to a single model family. Across the four heterogeneous architectures evaluated, accuracy increased by an average of 8.75 percentage points and precision by an average of 13.0 percentage points, with positive uplifts for every model. F1 score improved for three of four models, the only exception being DeepSeek-R1 whose weak baseline reasoning could not be fully compensated by the agent scaffolding alone. These cross-model patterns suggest that the framework generalises beyond the idiosyncrasies of any single LLM, an observation consistent with the empirical findings of Ke et al. (2025) for retrieval-augmented LLMs in surgical fitness assessment, and complementary to the broader trend toward standardised evaluation protocols in AI research (Yang et al., 2025; Zhang & Lu, 2021).

5. Discussion

5.1 Interpretation of Empirical Findings

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 screening applications 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 (+17.90 percentage points), F1 (+12.86 percentage points) and accuracy (+17 percentage points). 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 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.

5.2 Interpretability and Translational Implications

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 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. Translational deployment in primary care, telepsychiatry or community mental-health settings is therefore plausible, provided that appropriate human oversight, governance and validation procedures are in place.

5.3 Limitations

Several limitations should be acknowledged. First, the empirical 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 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 Comparison with Alternative Paradigms

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. The framework examined here 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 (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: this approach is simpler than the two-stage agent design 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.

5.5 Ethical, Regulatory and Engineering Considerations

Mental-health applications of AI raise distinctive ethical and regulatory questions that go beyond accuracy metrics. 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. 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. From a biomedical engineering perspective, deployment recommendations include local on-premise inference to address data sovereignty, strict logging policies, periodic re-

evaluation against updated benchmarks and integration with electronic health records using FHIR-aware agents (Mendoza et al., 2025). These engineering practices align with broader trends in cyber-physical system design (Lu, 2017) and Internet-of-Things-based healthcare infrastructure (Lu & Xu, 2019).

5.6 Future Research Directions

Three directions appear especially promising. First, multi-site clinical interview corpora should be assembled to corroborate the present findings under more rigorous labelling regimes and across cultures. Second, hybrid retrievers and adaptive top-k strategies should be explored to improve recall without sacrificing precision. Third, human-in-the-loop integration patterns that allow clinicians to query, contest and refine the model's evidence chains in real time may pave the way for genuine collaboration between AI and clinicians in mental-health practice. Looking further ahead, the convergence of RAG-LLMs with adjacent technologies, including blockchain-based audit trails (Lu, 2019; Yang et al., 2025), distributed-ledger informed-consent systems and on-device personal LLM agents, suggests a future in which evidence-grounded psychiatric AI is deeply embedded into health information infrastructure.

6. Conclusion

This article surveyed emerging trends in retrieval-augmented Large Language Models for mental health assessment and instantiated a representative two-stage agent-orchestrated framework. Across four heterogeneous open-weight LLMs at the 4–8B parameter scale, the framework increased accuracy by up to 17 percentage points and precision by up to 17 percentage points relative to direct prompting, with the largest improvements observed for the weakest baseline model. Equally importantly, the framework produces citation-supported outputs that meet the explainability requirements of clinical AI. By aligning language-model reasoning with verifiable, guideline-anchored evidence, retrieval-augmented LLMs combined with structured agent reasoning constitute a viable, scalable and clinically interpretable pathway for AI-assisted mental health screening, with practical implications for translational research, biomedical engineering and primary-care practice.

Ethics approval and consent to participate

Not applicable. This study used a publicly available dataset of de-identified social-media posts and did not involve human or animal subjects.

Consent for publication

Not applicable.

Availability of data and materials

The mangoesai/DepressionDetection dataset used in this study is publicly available on Hugging Face. The system prompts and analysis scripts are available from the corresponding author upon reasonable request.

Funding

The authors declare that no external funding was received for this work.

Competing interests

The authors declare no competing interests.

AI use disclosure

Generative AI tools were used solely for English-language polishing and citation formatting. All scientific content, analyses and conclusions are the responsibility of the human authors.

Author contributions

H. Liu and B. Chen: Conceptualization, Methodology, Writing - Original Draft; X. Zhang: Clinical context and data interpretation, Writing - Review & Editing; M. Sun: Software, Formal Analysis; J. Wang: Validation, Investigation; Y. Pan: Visualization, Writing - Review & Editing; B. Chen: Supervision, Writing - Review & Editing.

Acknowledgements

The authors thank colleagues in the affiliated departments for valuable feedback on early drafts, and the open-source LLM community for releasing the model weights used in this study.

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., Montana-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

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

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

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., et al. (2023). The future landscape of large language models in medicine. Communications Medicine, 3, 141. https://doi.org/10.1038/s43856-023-00370-1

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

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

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

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., et al. (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

Gomez-Gomez, I., Benitez, I., Bellon, J., Moreno-Peral, P., Olivan-Blazquez, B., Claveria, 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

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

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

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W. T., Rocktaschel, 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

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

Lu, Y. (2017). Cyber physical system (CPS)-based Industry 4.0: A survey. Journal of Industrial Integration and Management, 2(3), 1750014. https://doi.org/10.1142/S2424862217500142

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

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. (2024). Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Lu, Y., Pisarenko, Z. V., Yang, L., & Ye, C. (2024). Advancing decision-making: The role of management analytics in modern business practices. Nanotechnologies in Construction, 16(5), 431-440. https://doi.org/10.15828/2075-8545-2024-16-5-431-440

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Lu, Y., Zheng, X., Li, L., & Xu, L. D. (2020). Pricing the cloud: A QoS-based auction approach. Enterprise Information Systems, 14(3), 334-351. https://doi.org/10.1080/17517575.2019.1669827

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

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

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

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

Schick, T., Dwivedi-Yu, J., Dessi, 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

Shi, X., Lee, S., & Kim, M. (2024). ArgMed-Agents: Explainable clinical decision reasoning with LLM discussion 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

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

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., Ruckle, 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

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

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

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Roziere, 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

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & 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

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

Vrdoljak, J., Boban, Z., Vilovic, M., Kumric, M., & Bozic, 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

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

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

Yang, L., Hou, Q., Zhu, X., Lu, Y., & Xu, L. D. (2025). Potential of large language models in blockchain-based supply chain finance. Enterprise Information Systems, 19(11), 2541199. https://doi.org/10.1080/17517575.2025.2541199

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

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

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

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