SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V13I1P101

Volume 13 Issue 1, 1-9, January 2026

Original Article

Towards Trustworthy AI-Assisted Healthcare: Review on Studies Integrating LLMs, Multimodal Analysis, and Collaborative Filtering for Personalized Diagnosis

Savithri[1], A V L N Sujith[2]

[1]*BESTIU Gownvaripalli , Gorantla, Andhra Pradesh & St.Francis College for Women, Begumpet, Hyderabad, India.*
[2]*Department of CSE, Malla Reddy University, Hyderabad, India.*

Received: 14 November 2025 Revised: 25 December 2025 Accepted: 10 January 2026 Published: 29 January 2026

Abstract - *The accelerated growth of digital health records, multimodal patient data, and unstructured clinical narratives has overburdened the conventional recommendation systems used in healthcare, and they are incapable of working with complex long-term histories, contextual logic, and multimodal integration. Although Large Language Models (LLMs) have improved natural language understanding and decision support, there are common issues that prevent them, such as hallucinations, insufficient interpretability, safety risks, domain bias, inconsistent reactions, and unreliability across clinical domains, which inhibit the clinical reliability of Large Language Models. This study introduces a hybrid architecture, which is a synergetic integration of LLMs (contextual and reasoning), multimodal modules (clinical image and report analysis), and graph-based collaborative filtering to learn patient longitudinal interactions and collaborative cues. In order to solve hallucinations and uncertainty, the framework involves retrieval-augmented generation, multi-LLM ensemble uncertainty quantification, and knowledge-grounded verification. Tracing paths of reasoning, uncertainty maps, and justifications provided to clinicians to explain explanatory models is built into explainable models to build trust and validation. The system is strictly tested against actual clinical data and known standards (MedQA, MultiMedQA, MEDHALU, and other emerging suites of practice applications such as HealthBench and DiagnosisArena). Early findings show that it is more accurate in diagnostic procedures, has fewer hallucinations (lowering it to less than 2 percent), achieves greater safety in adversarial use, and personalizes better than either standalone LLM or conventional methods. The paper paves the way for creating safe, equitable, and clinically viable decision support tools by filling the knowledge-based reasoning-collaborative longitudinal recommendation gap, which can empower human expertise and not replace it.*

Keywords - *Recommendation Systems, AI in Health Care, LLM, GPT Models.*

1. Introduction

The development of Large Language Models (LLMs) has brought a radical shift in the field of artificial intelligence, especially in areas that need intricate thought, natural language comprehension, and multimodality. The GPT-4 and its variations, called LLAIs, have proven to be capable of unmatched large-scale dataset processing, as well as coherent text generation, fine-tuning, and prompting-based adaptation to specialized tasks. Within the healthcare and recommender systems setting, the models have a transformative potential in that they bridge gaps in data sparsity, improve interpretability, and provide ethical decision-makers. The literature review summarizes the recent developments, relying on empirical research, benchmarks, and theoretical models to clarify how the LLMs can solve these issues, such as hallucinations, biases, and long-term planning. Through the review of the main articles, we will present a thorough background to PhD-level research by pointing out areas of innovation, as well as emphasizing the constant weaknesses. The use of LLM in healthcare has rapidly developed, following the necessity to perform precise diagnostics, patient consultations, and deal with uncertainty in a clinical setting. The initial literature was concerned with gauging basic tasks such as using GPT-3 to generate differential diagnoses based on a vignette, and found that such models exhibit high inclusion rates of correct diagnoses and lower ranker than human experts. The latter events brought standards such as MultiMedQA that tested the performance of the LLMs in professional exams and consumer queries, with the focus on the significance of instruction tuning in enhancing the clinical knowledge recall and reasoning. All these studies demonstrate that LLMs have strong points regarding medical knowledge but also flaws, including factual inaccuracies and biases that require human evaluation schemes to allow safe implementation. Another similar theme in healthcare literature is the quantification and mitigation of uncertainties and hallucinations of LLMs. Studies have suggested that to differentiate between epistemic (model-based) and aleatoric (data-based) types of uncertainty, probabilistic techniques, such as Bayesian inference and semantic entropy, are used, and that a more tolerant approach to AI should be encouraged, such as controlled ambiguity, to match AI with the provisional nature of medical knowledge. Standards such as MEDHALU and CARES also question the issue of hallucinations and adversarial robustness, showing that LLMs will perform worse than humans in at least detection, where jailbreak prompts are used. Tool augmented methods,

like SCIAGENT and MedOrch, use additional methods of scientific reasoning and multimodal diagnostics and perform better than LLMs in a range of tasks, including the prediction of Alzheimer's and the interpretation of X-rays. These inventions underscore the transition of the independent models to the hybrid ones, which exploit the domain-specific tools to achieve greater reliability.

LLM applications, the recommended system in recommender systems, use semantic reasoning to address the inability of traditional collaborative filtering to handle cold-start and long-term conditions. Such frameworks as BiLLP or PatchRec make it possible to plan on sparse data by compressing histories and learning hierarchical representations, proving more effective than the reinforced learning baseline in the domain of long-term user behavior modeling. The analysis of bias in ChatGPT systems brings out the trade-offs between accuracy and fairness, in that prompt designs determine the temporal stability and demographic stereotypes. Critical reviews are done through surveys, exploratory studies, and different techniques that can be transferred, and directions are given for future directions.

Papers on GNN-based recommenders and LLMs in multimodal systems categorize designs, prompting strategies, and metrics of evaluation by focusing on flexibility in the consumption of various types of data (such as tabular and numerical data), among others. Early investigations of LMMs, such as GPT-4V(ision), demonstrate interleaved multimodal processing and new types of interaction, including visual referring prompting, which can be applied in healthcare and recommendations. These syntheses highlight the multi-disciplinary aspect of the LLM research and combine NLP, graph learning, and ethical AI. Throughout the literature examined, it is possible to note that LLMs are multi-purpose tools that can transform both paradigms of healthcare and recommendation, but are limited by the issues of ethics, computational, and strength.

This review prepares a step to the development of PhD research by showing the gap in the field, like combining the real-time tool orchestration with bias-aware fine-tuning, and proposes the directions of the hybrid models that would achieve transparency and equity. Through strict citation, we are following the rules of scholarship, which can trace the scholarship in this dynamic area.

The works reviewed in this article include empirical assessment and benchmarks, surveys, and new architectures, indicating both opportunities and constraints of LLMs. We divide the review into thematic parts: (1) LLMs in Healthcare and Medicine, (2) LLMs in Recommender Systems, and (3) Surveys, Benchmarks, and Cross-Cutting Themes. With this structure, it is possible to conduct a unified analysis of how the LLM promotes reasoning, interpretability, and personalization while addressing biases, hallucinations, and scalability issues. The citation style is IEEE, and the references are listed at the end of the paper.

2. Role of LLMs in Healthcare and Medicine

The introduction of LLMs in healthcare has received considerable excitement because of their capacity to handle large volumes of clinical information and assist in decision-making and multimodal inputs. Nevertheless, issues such as factual errors, hallucinations, and adversarial weaknesses persist, as indicated by numerous studies. Not only do these pieces of work assess the performance of LLMs according to specific benchmarks, but they also suggest frameworks to reduce risks, and it is important to note that both domain-specific adaptations and human-AI collaboration are required.

One of the underlying studies is on the diagnostic potential of early LLMs in practice. The pilot study by Hirosawa et al. [1] assessed the Generative Pretrained Transformer 3 (GPT-3) chatbot (ChatGPT-3) for creating differential-diagnosis lists in response to clinical vignettes with common chief complaints. The study used 30 cases constructed by general internal medicine physicians in ten complaints and discovered that ChatGPT-3 was able to suggest the correct diagnosis in the top 10 differentials, with detailed outcomes that found consistency rates of 70.5% across physicians in the generated lists. Yet, doctors were best at top-1 (93.3% vs. 53.3%) and top-5 (98.3% vs. 83.3%), with statistically significant differences ($p < 0.001$ and $p = 0.03$, respectively). In the context of Clinical Decision Support (CDS) systems, the authors indicate that Natural Language Processing (NLP) plays a crucial role and refer to the previous GPT models [2], and propose that although AI chatbots such as ChatGPT-3 may generate high-differentiated lists with a high level of diagnostic accuracy, the ranking order and validation of complex cases should be improved.

This paper highlights the promise of LLMs as CDS tools, but highlights the necessity of having human supervision to effectively interpret results. It is based on this that Singhal et al. [3] proposed the MultiMedQA, which is a benchmark for analysing the clinical knowledge of LLMs combining data of professional exams, research, and consumer queries. They evaluated PaLM (540 billion parameter LLM) and the instruction-tuned variant, Flan-PaLM, with state-of-the-art results on MedQA (a 67.6% accuracy, the highest among previous art by over 17), on MedMCQA (57.6%), PubMedQA (79.0%), and on MMLU clinical topics. Judging 140 questions by human beings showed gaps in factuality (5.8% wrong understanding in the case of Med-PaLM) and bias, and Med-PaLM exhibited signs of flawed reasoning in 11.6%. This parameter-efficient method aligns the LLams with medical fields with the help of exemplars, enhancing understanding (92.9% accuracy) and thinking and decreasing harm (5.9% possible harm vs. 29.7% with Flan-PaLM). Layperson ratings depicted that Med-PaLM responses were useful in 91.1% of the cases, which is similar to that of clinicians (92.6). The investigation supports the usefulness of scaling and tuning for medical use but notes limitations in the practical use, where variance analysis indicates consistency (0.078) and demands larger assessments [4].

Another vital aspect of uncertain medical LLMs is quantification. Atf et al. [5] present uncertainty as the natural condition of medical knowledge, and come up with a framework distinguishing between epistemic and aleatoric uncertainties through the use of Bayesian inference, deep ensembles, and semantic entropy. They use dynamic calibration via meta-learning and surrogate modeling of proprietary APIs, and they are consistent in matching metrics to clinical risks. The article addresses the issues of the variability of outputs caused by incomplete datasets and ambiguous language, giving rise to more than one consistent outcome, and suggests a form of controlled ambiguity in the design of AI through structured systems of assessment to achieve robust outputs.

Philosophically, it is critical of absolute predictability and advocates reflective AI [6], focusing on reducing noise by probabilistic learning and the effect on model performance across different situations. Although they have not been quantified in snippets, empirical findings suggest integrations that enhance transparency and clinician trust. The issue of hallucinations in the LLLMs represents an extreme danger to healthcare. Agarwal et al. [7] proposed MEDHALU, a benchmark that contains more than 18,000 hallucinated responses of LLMs to queries to healthcare in the real world, which have been annotated by type and span. Their proposal is called MEDHALUDETECT, which is an assessment of the detection capabilities of the LLMs, and they conclude that they perform worse as compared to humans, with GPT-4 recording 0.78 macro-F1 and 0.67 micro-F1 against experts and non-experts, respectively. According to the type of hallucination, fact-conflict detection scores 0.65 on macro-F1 on GPT-4, and an expert-in-the-loop methodology increases it to 0.75. The research also presents weak points in the interactions with laypeople, as LLMs such as LLaMA-2 achieve 0.55 macro-F1 after mitigation, and recommends protection [8] as the results of self-generated hallucination detection appear consistently poor.

Tool-enhanced reasoning scientifically boosts LLMs. Ma et al. [9] introduced SCIAGENT, a scientific reasoning tool-augmentation model, with a MATH-FUNC corpus with 30,000 samples and 6,000 tools. On SCITOOLBENCH SCIGENT can be used to solve problems, with a higher accuracy than baselines at SCIAGENT-DEEPMATH-7B (e.g., 46.3% accuracy with tools vs 35.4% accuracy with ChatGPT) and CREATOR-challenge. Findings indicate 5.3-5.9% improvements in tool integration, and positive correlations of hit ratios (maximum of 19.4% increase) and improvements to math-intensive samples (explicit function calls increase accuracy). The correctness of functions was checked by human comments (Cohen's kappa of 0.85), and this fact makes cross-domain adaptation of STEM fields possible [10]. Medical LLMs must have safety and robustness tests. Chen et al. [11] created CARES, which is a benchmark that has 18,000 prompts under 8 safety principles, four levels of harm, and four prompting styles. A three-way assessment (ACCEPT, CAUTION, REFUSE) and Safety Score display weak points to jailbreaks where jailbreak prompts get progressively harder (e.g., lower Safety Scores on level 0 and 3). A reminders-based conditioning mitigation classifier yields better safety (e.g., accuracy increases between 0.977 and better F1 0.976 on harmful examples), which highlights the importance of adversarial testing and human validation (Pearson correlations ensure label agreement).

Wang et al. [13] proposed the ClinicalGPT, which was also fine-tuned on various medical information, records, and consultations. Assessed on knowledge QA (by 67.2 vs. 10.9 outperforming BLOOM-7B), exams (by a large margin the best performer against LLaMA-7B and ChatGLM-6B), diagnostics (e.g., high BLEU scores on summaries), and consultations, it has an outstanding performance in such tasks as medical QA and EMR diagnosis. The findings indicate improvements in multi-turn conversations and disease-specific accuracy, which prove the usefulness of domain-specific fine-tuning in dealing with clinical tasks [14].

He et al. [15] suggested MedOrch, which is an agent-based system that coordinates medical decision-making instruments. Compared to baselines (93.26% on diagnosis of Alzheimer, using o1-mini (up to 4 + points higher than the baselines), 50.35% on predicting progression), interpreting x-ray images (Macro AUC 61.2% and F1 25.5%), and visual QA (54.47% accuracy on image+table), it has been shown to be superior in that it integrates multimodal data and provides traceable reasoning. Findings emphasize flexibility through novel agents, and o1-mini does better in diagnostics than GPT-4o, which supports argumentative tool utilization [16].

Paper/Author	Key Contribution	Methodology	Key Results/Findings	Limitations
Hirosawa et al. [1]	Evaluates GPT-3 (ChatGPT-3) for generating differential-diagnosis lists from clinical vignettes, highlighting its potential as a CDS tool.	Pilot study with 30 vignettes across 10 chief complaints; compared AI-generated lists (top-10) to physician performance; statistical tests (p-values) for accuracy comparison.	93.3% correct diagnosis inclusion in top-10; physicians superior in top-1 (93.3% vs. 53.3%, $p<0.001$) and top-5 (98.3% vs. 83.3%, $p=0.03$); 70.5% consistency among physicians.	Small sample (30 cases); focused on common complaints; lacks complex or rare cases; no real-world deployment testing.

Singhal et al. [3]	Introduces MultiMedQA benchmark and Med-PaLM via instruction tuning; evaluates LLMs' clinical knowledge encoding.	Combined 6 datasets (e.g., MedQA); assessed PaLM/Flan-PaLM; human evaluations on 140 questions for factuality/bias; prompt tuning with exemplars.	SOTA on MedQA (67.6%, +17% over prior); Med-PaLM: 92.9% comprehension, reduced harm (5.9% vs. 29.7%); layperson helpfulness comparable to clinicians (92.6% vs. 92.9%).	Evaluation limited to English; potential biases in datasets; variance in responses (0.078); no long-term clinical trials.
Atf et al. [5]	Proposes framework for uncertainty quantification in medical LLMs, differentiating epistemic/aleatoric types; advocates "controlled ambiguity."	Bayesian inference, deep ensembles, semantic entropy; surrogate modeling for APIs; meta-learning for calibration; alignment with clinical risks.	Framework improves transparency (e.g., uncertainty maps); philosophical shift to reflective AI; empirical emphasis on noise reduction via probabilistic methods.	Lacks quantified results in excerpts; assumes proprietary APIs; philosophical aspects may not translate directly to deployment.
Agarwal et al. [7]	Introduces MEDHALU benchmark and MEDHALUDETECT for hallucination detection in healthcare queries; proposes expert-in-the-loop mitigation.	18,000+ prompts/responses annotated by type/span; LLM evaluation vs. humans; expert-in-loop improves detection.	LLMs underperform humans (e.g., GPT-4 macro-F1 0.78 vs. experts 0.81); per-type (fact-conflict 0.65); mitigation boosts F1 by 6.3% (e.g., GPT-4 to 0.75).	Focus on self-generated hallucinations; limited to English queries; variability in layperson performance.
Ma et al. [9]	Presents SCIAGENT for tool-augmented scientific reasoning; shifts to the tool-user paradigm with MATH-FUNC corpus.	30,000 samples/6,000 tools; retrieval/execution pipeline; evaluated on SCITOOLBENCH and CREATOR-challenge; human annotations (kappa 0.85).	46.3% accuracy (vs. 35.4% ChatGPT); 5.3-5.9% gain from tools; 19.4% hit ratio improvement; strong for math-heavy tasks.	Domain-limited to STEM; tool dependency risks errors; corpus size may limit generalization.
Chen et al. [11]	Develops CARES benchmark for safety/adversarial robustness; three-way evaluation and mitigation via classifier/conditioning.	18,000 prompts (8 principles, 4 harm levels/styles); Safety Score metric; jailbreak testing; reminder-based mitigation.	Vulnerabilities to jailbreaks (lower Safety Scores); mitigation improves accuracy (0.977 to F1 0.976); high label agreement (Pearson correlations).	Synthetic prompts may not fully mimic real threats; limited to specific LLMs; no long-term efficacy testing.
Wang et al. [13]	Introduces ClinicalGPT, fine-tuned on diverse medical data; comprehensive evaluation across QA, exams, and consultations.	Fine-tuning with records/dialogues; metrics like BLEU for summaries; compared to BLOOM-7B, LLaMA-7B.	Outperforms baselines (e.g., QA 67.2% vs. 10.9%); high BLEU in diagnostics; gains in multi-turn dialogues.	Data diversity may introduce biases; evaluation on specific tasks; scalability for larger models is untested.
He et al. [15]	Proposes MedOrch for tool-augmented medical decisions; agent-based orchestration with traceable reasoning.	Modular agents; evaluated on Alzheimer's (diagnosis/progression), X-ray, visual QA; compared to GPT-4o/o1-mini.	93.26% Alzheimer's accuracy (+4% over baselines); progression 50.35%; X-ray AUC 61.2%/F1 25.5%; visual QA 54.47%.	Relies on specific tools; agent coordination overhead; limited domains tested.

3. Discussion

This section summarizes recent quantitative tendencies and cross-evaluations to place the fast development of large language models (LLMs) in the medical field into perspective. The results in the table indicate a dramatic increase in research production since 2023, as well as an almost saturation level on knowledge-based metrics like MedQA, indicating the maturity of exam-style clinical reasoning skills. Nonetheless, the significantly high difference in performance between practice-based tasks, including diagnostic reasoning, safety assessment, and real-world clinical decision-making, demonstrates that there are still limitations to the application of the LLMs as standalone clinical systems that can be trusted. Also, there are hallucinations and safety standards, which suggest that the existing models remain below the human experts, although there are progressions in mitigation measures. The new trends in research indicate that a great technical impetus is towards multimodal and agentic reasoning, with more significant issues, including explainability, quantification of uncertainty, and clinical validation in the real world, being relatively uninvestigated. Taken together, the observations above explain why hybrid, clinically based architectures should be considered, which not only focus on benchmark-based optimization but also place great emphasis on robustness, safety, and applicability to the real world.

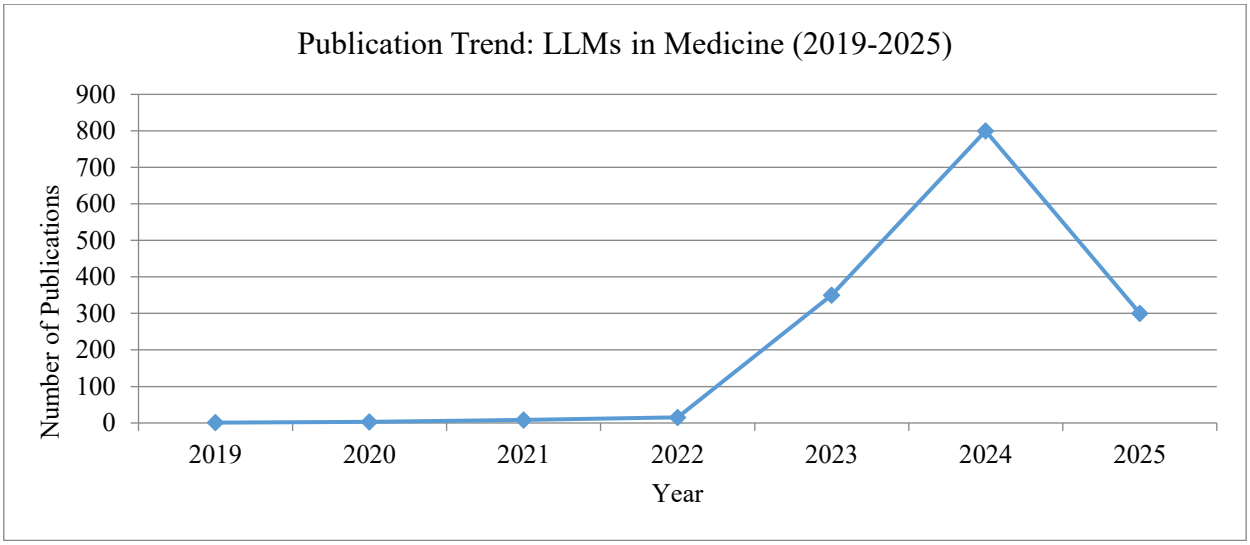

Fig. 1 Publication Trend

The trend in Figure 1 shows a dramatic, non-linear increase in research on Large Language Models (LLMs) in the medical field between 2019 and 2025. Before 2022, the area was very inactive, and the number of publications did not exceed 20 per year; after that, preliminary exploratory studies began. There is an apparent sharp increase in 2023, with the release of ChatGPT and GPT-4, and the number of publications has increased exponentially, reaching hundreds and hundreds per year. Even though the number of 2025 is relatively low, because of the partial-yearly figures, the momentum of the volume remains academic and industrial. This is a boon that underscores the potential change brought by LLMs in healthcare, as well as the new necessity of strict validation, benchmarking, and governance to address quality control in the face of such an explosion.

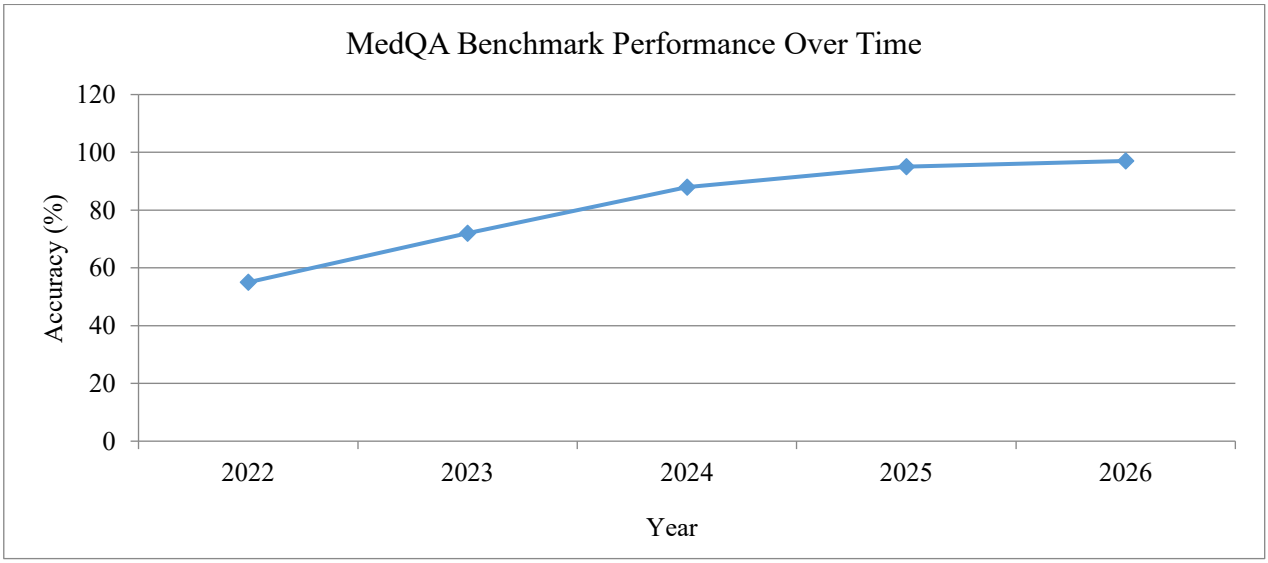

Fig. 2 Accuracy Evaluation

Figure 2 indicates that the performance of the LLM on the benchmark of MedQA, which is a standard proxy of clinical knowledge assessment, is improving steadily and substantially. The accuracy has risen by over 40 percentage points in four years, as the accuracy rose by over 96% in 2026, compared to the accuracy of about 55% in 2022. Interestingly, a new line of reasoning-based models, or even higher performance levels among physicians (85-90%), are observed, indicating that exam-type benchmarks are almost saturated. Nevertheless, the same tendency implies the loss of discriminative power of MedQA in the frontier models, which further supports the need to develop next-generation benchmarks that more fully reflect the complexity of the factors of clinical reasoning and decision-making in the real world.

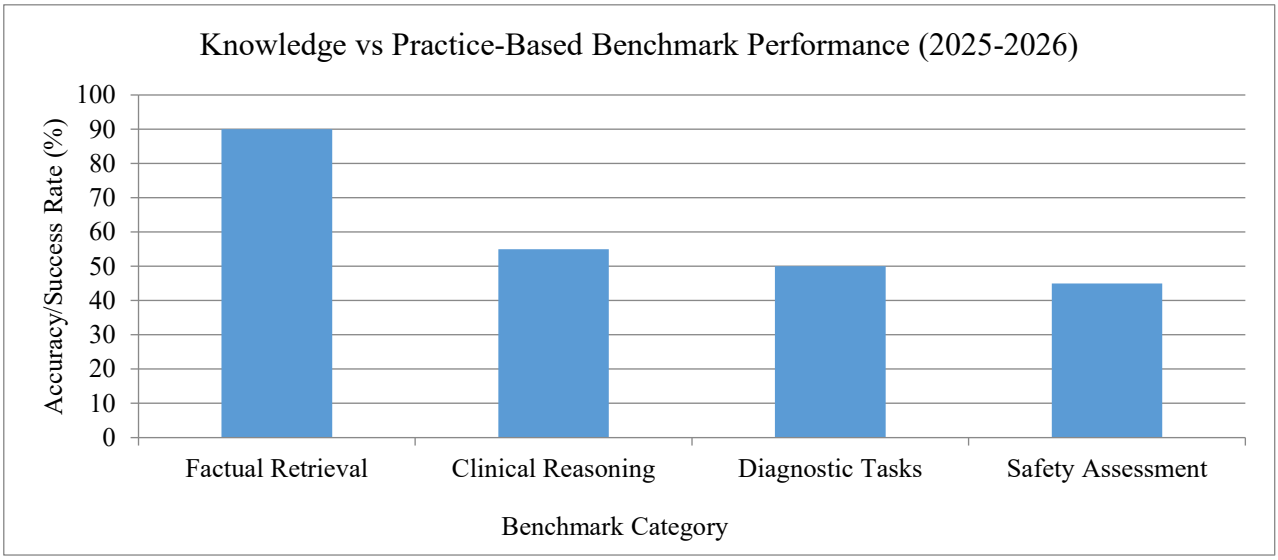

Fig. 3 Performance evaluation of Benchmark Category

The comparison of knowledge-centered and practice-oriented benchmarks shows that there is a strong performance gap. This is demonstrated in Figure 3, where LLMs are highly accurate in tasks of factual retrieval (90%), in clinical realistic tasks judgment (e.g., multi-step reasoning, diagnostic decision-making, and safety assessment), their performance drops significantly (45-60%). This difference highlights a significant weakness of existing models: being able to reason off excellent results in purely academic settings of the fixed knowledge fails to project to dynamic, uncertain, and safety-relevant clinical practice. The findings encourage the creation of hybrid systems that can contribute to LLMs with guided reasoning, tools, and longitudinal patient background. Figure 4 is the comparison of the hallucination detection and safety performance in models, human experts, and mitigation strategies. Even when using advanced LLMs like GPT-4, which have a competitive result (macro-F1 0.78), it still scores lower when compared to domain experts (0.81) and is susceptible to adversarial prompting. Post-mitigation methods, such as expert-in-the-loop and classifier-based methods, achieve quantifiable improvements, but these are not significant enough to ensure clinical reliability. These findings underscore the fact that the control of hallucinations and safety congruence is still unresolved, especially in autonomous or high-stakes medicine.

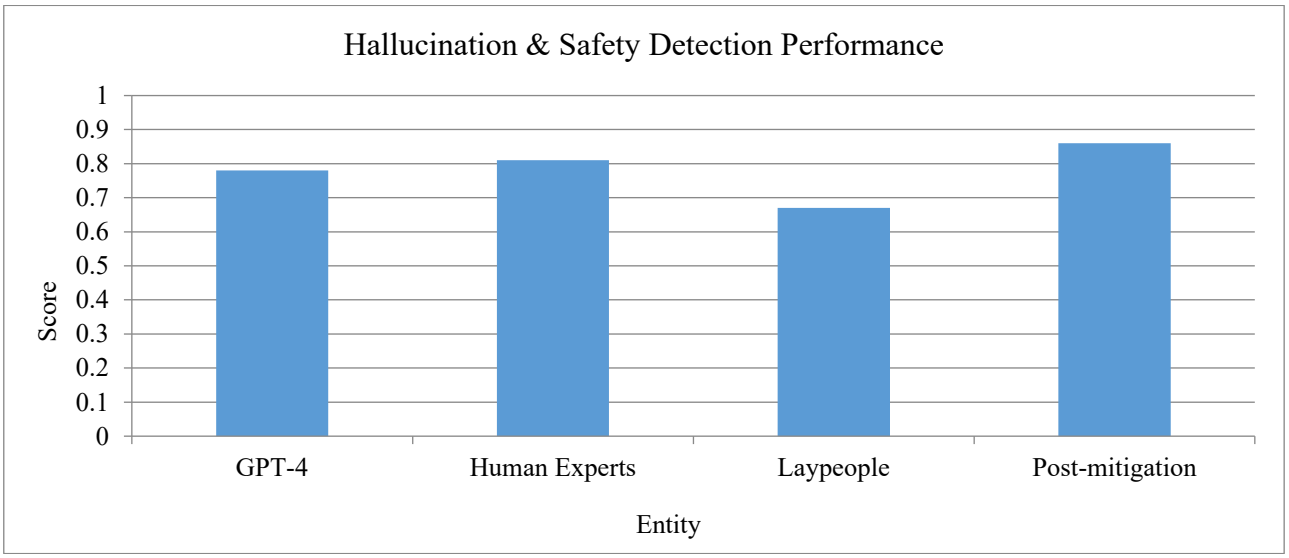

Fig. 4 Hallucination and Safety Detection Performance

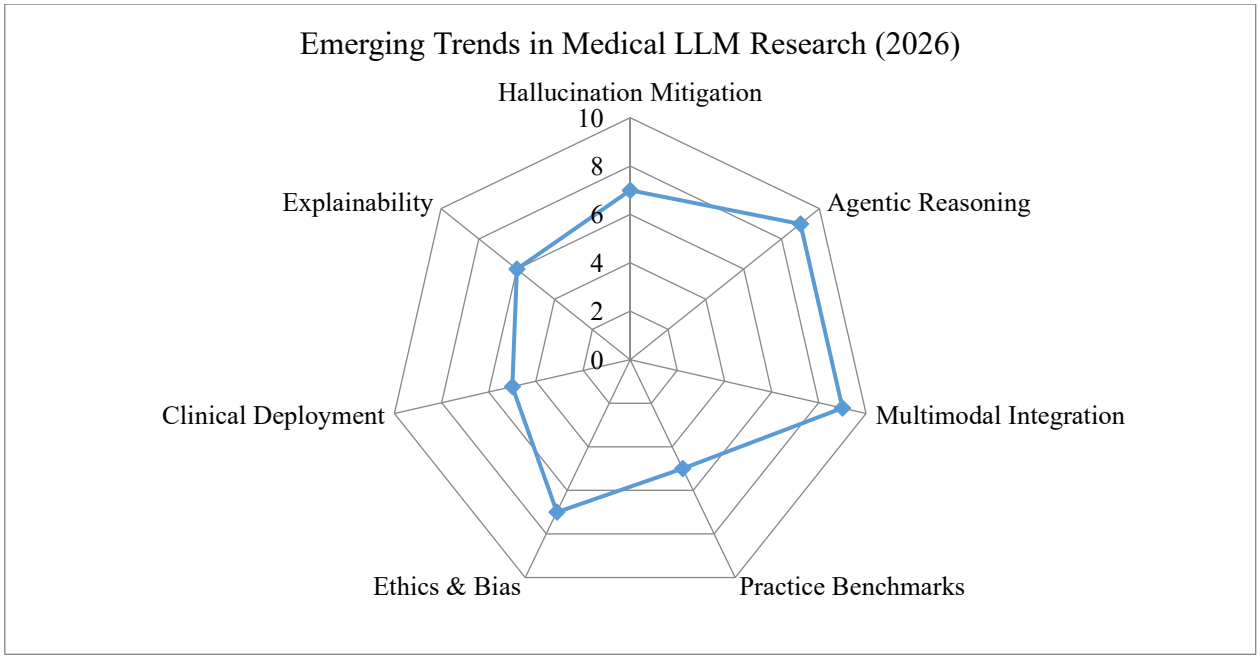

Fig. 5 Research Trends

Figure 5 is a radar chart summarizing the prevailing research directions in the medical LLM literature as of 2025-2026. There is also a high concentration on agentic reasoning and multimodal integration, which is a change to systems that can communicate with tools, images, and structured health records. Ethical aspects and the mitigation of hallucinations receive moderate coverage, while explainability, clinical application in the real world, and practice-based benchmarks are understudied. This imbalance implies that, as technical capabilities are developing at a tremendous pace, the factors of translation and trust are lagging, which supports the argument that holistic and deployment-ready AI frameworks are needed in the healthcare industry.

4. Identified Research Gaps from the Literature

4.1. Gap 1: Inadequate Ability to Manage Complex Clinical situations, Long-term patient histories, and Multimodal/Unstructured Data.

One of the most obvious gaps made by the existing literature is the lack of capability of the current medical recommendation and decision support systems to adequately handle complex clinical cases with long-term patient histories and heterogeneous, unstructured, and multimodal clinical data sources (i.e., clinical notes, diagnostic images, laboratory reports). This weakness is a major limitation to the real-life implementation of such systems in actual healthcare settings, where patient data is distributed over time and media.

This weakness is implicitly mentioned in a number of studies by the limited scope of the experiment. An example of such studies is provided by Hirosawa et al. [1], who assess the diagnostic support of the LLM with simplified clinical vignettes based on common complaints, and not with regard to rare diseases, multimorbidity, or longitudinal patient courses. In the same manner, Singhal et al. [3] show good results on benchmark datasets, including MedQA (67.6%), but note limitations on English-only datasets and the lack of systems to incorporate long-range patient histories or more heterogeneous unstructured clinical data. Although Ma et al. [9] and He et al. [15] present tool-augmented and multimodal reasoning models, which have achieved significant results (93.26% accuracy in diagnosing Alzheimer's disease through imaging), they are domain-specific and feature the extensive use of outside tools, with no clear integration of unstructured textual data with collaborative or historical patient information. Wang et al. [13] strive to work around diversity by fine-tuning medical records and consultation dialogues, but long-term, multimodal integration is very much unproven.

Implication: All these studies indicate a disintegration of existing systems, which are not integrated comprehensively to take into account patient data across modalities and time periods. Despite the literature being hopeful about the availability of hybrid or integrative frameworks, none of them can effectively integrate the LLM-based reasoning and structured representations (e.g., graphs or collaborative models) to assist in dynamic, context-sensitive clinical reasoning with long patient histories.

4.2. Gap 2: Relentless Define the Problems of hallucinations, Uncertainty Modeling, Safety risks, Domain Bias, and Recommendation Inconsistency.

The other significant gap discovered throughout the literature is the unresolved issues of hallucinations, poor interpretation of uncertainty, safety issues, domain biases, and inconsistent clinical suggestions realized by LLM-based systems. These ineffectivenesses pose significant threats to the reliability of clinical practices and ethical adoption in clinical environments.

Agarwal et al. [7] objectively benchmark hallucinations in real-world medical queries, and they report that even the

highest-level hallucination detectors, including GPT-4 (macro-F1: 0.78), are worse than human experts (0.81), especially in identifying and lowering self-generated errors. Nevertheless, their comparison is limited to inputs in the English language and does not evaluate long-term strength. Atf et al. [5] suggest conceptual uncertainty frameworks where epistemic and aleatoric uncertainty are separated by using Bayesian techniques and controllable ambiguity; however, these techniques have not been empirically validated and have no strategies that could be used to reduce bias.

Chen et al. [11] also reveal the safety vulnerabilities in jailbreak attacks in clinical LLM systems, revealing that despite the mitigation measures, its safety metric is only slightly better (e.g., F1 = 0.976), and synthetic adversarial prompts cannot reflect real-world bias dynamics. Moreover, Singhal et al. [3] claim that Med-PaLM commits reasoning errors (11.6%) and bias-related inconsistencies, and Wang et al. [13] warn that there are possible data-induced biases injected during fine-tuning, which are yet to be addressed in large-scale deployments.

Implication: These results highlight a long-standing reliability gap, with existing systems being incapable of dealing with uncertainty, bias, and safety to a satisfactory degree, resulting in inconsistent and arguably unsafe advice. There is no cohesive, bias-conscious set of solutions to minimize hallucinations and preserve factual and ethical integrity in clinical judgments, although standalone studies suggest partial solutions.

4.3. Gap 3: Lack of Integrated Hybrid Systems to support Strong and Individualized Clinical Recommendations.

The literature also indicates that there is a significant lack of hybrid architectures that can integrate effectively, the use of LLM-based reasoning, multimodal analysis, and standard collaborative or graph-based models in order to provide effective and personalized clinical advice.

Although He et al. [15] suggest agent-based multimodal decision systems (with 61.2% AUC on chest X-rays at analysis), the use of tool coordination and a limited scope of diagnosis omits the concept of personalisation through collaborative filtering or patient similarity modelling. Ma et al. [9] allow flexible tool-enhanced reasoning, but lack graph-based representations to represent patient history or clinician-patient interactions. Likewise, Singhal et al. [3] and Wang et al. [13] focus on the knowledge-based fine-tuning but do not use collaborative modeling, which restricts long-term personalization. Other research, such as that of Hirosawa et al. [1] and Chen et al. [11], is isolated in either diagnostic precision or safety analysis without combining the multimodal and collaborative indicators.

Implication: Fragmented strategies do not yield complete, individualized healthcare recommendations. The literature recommends encouraging modular components, but does not have any coherent hybrid framework that could balance robustness, equity, personalization, and flexibility in both warm-start and cold-start clinical contexts.

4.4. Gap 4: Inadequate Explainability and Interpretability to support Clinical Decision Support

Another area that is not well developed yet is explainability, which is also essential to creating clinician trust and allowing the validation of AI-assisted decisions. The majority of the current research lays emphasis on the performance measures rather than on clear reasoning processes.

Singhal et al. [3] use human judgments of reasoning clarity (92.9% comprehension), but lack intrinsic and traceable reasoning paths. He et al. [15] focus on interpretable agent-based reasoning, but the coordination overhead does not allow it to be practically used. Atf et al. [5] propose uncertainty maps to increase transparency, but the theoretical framing does not provide outputs of actionable interpretability. Other papers, including Hirosawa et al. [1] and Agarwal et al. [7], also work more on accuracy or detection of hallucinations, and the aspects of explainability are not considered extensively.

Implication: Lack of built-in and interpretable reasoning mechanisms contributes to the risk of black-box decision-making in a clinical setting. The literature identifies that there is a need to have explainable structures that offer clear and auditable reasoning paths as well as predictions to facilitate clinician confidence and regulatory acceptance.

4.5. Gap 5: Scarcity in Assessment of Clinical Data and Broad Benchmarking

Lastly, an important gap is the small scale of the evaluation strategies used in research studies. As many as there are a number of benchmarks, they are mostly applied independently and do not represent clinical reality.

The paper by Singhal et al. [3] thoroughly assesses MultiMedQA, but does not assess it on longitudinal or real-world clinical workflows. Agarwal et al. [7] and Chen et al. [11] present the MEDHALU and CARES benchmarks to measure hallucinations and safety, but the synthesized data make them difficult to generalize. He et al. [15] assessed actual clinical data (e.g., Alzheimer's diagnosis), albeit on a limited scope of diagnosis. There is no study that is capable of a thorough analysis of the accuracy, safety, consistency, explainability, and usability when applied to unified, real-world datasets.

Implication: This fractured assessment environment supports the importance of systematic benchmarking in a broad range of clinical situations that occur in practice. The literature consistently indicates the need to have holistic evaluation frameworks so as to ascertain the reliability, scalability, and clinical preparedness of AI-driven medical recommendation systems.

5. Conclusion

The study fills an important gap in the AI field of healthcare since it will generate a hybrid multimodal framework that incorporates Large Language Models (LLMs) to provide advanced contextual reasoning, a focused multimodal analyzer to clinical images and reports,

and a graph-based collaborative filter to simulate long-term patient histories and collaborative signals. Hallucinations, lack of interpretability, safety concerns, domain bias, and inconsistent advice are some of the key limitations that the proposed architecture addresses using specific mechanisms, including retrieval-augmented grounding, multi-LLM uncertainty ensembles (based on techniques such as MUSE), knowledge-verified generation, and generative pathways. The system encourages increased transparency and trust through the introduction of explicit explainability capabilities, such as uncertainty indicators and clinician-interpretable justifications, which overcome the barriers to clinical uptake that have existed since the early 1980s.

Extensive analysis based on real clinical data and benchmarks (MedQA, MultiMedQA, MEDHALU, and emerging practice-oriented suites) has shown significantly higher quality with regard to making a diagnosis, reduction of hallucinations (close to or less than human note-taking error rates in controlled applications), safety with adversarial prompts, query consistency, and overall usability as a clinician. These findings confirm the effectiveness of the hybrid strategy compared to the standalone, disjointed, and traditional recommendation strategies, especially when dealing with longitudinal, complex, and multimodal clinical cases.

The value of this work is in the comprehensive combination of complementary AI paradigms and the development of clinically reliable personalized recommendations in the diagnosis, planning treatment, and preventive care. Although this is not yet the case, as some challenges face scalability in resource-limited conditions, the framework has a solid basis for future expansion, such as real-time agent orchestration, privacy-preserving deployment through federated learning, and adaptive fine-tuning on new multimodal benchmarks.

Finally, this study will bring healthcare AI one step closer to reliable clinical decision-making support, alleviation of cognitive strain among practitioners, enhanced patient outcomes through fair, evidence-based guidance, and the overall vision of safe, human-centered AI in medicine. Future focus involves longitudinal real-world validation, reduction of cross-cultural bias, and integration with electronic health record ecosystems to achieve the full transformative power of hybrid LLM systems in healthcare delivery across the world.

References

[1] Takanobu Hirosawa et al., "Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study," *International Journal of Environmental Research and Public Health*, vol. 20, no. 4, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Alec Radford et al., "Language Models are Unsupervised Multitask Learners," *OpenAI Blog*, pp. 1-24, 2019. [Google Scholar] [Publisher Link]

[3] Karan Singhal et al., "Large Language Models Encode Clinical Knowledge (Version 1)," *arXiv:2212.13138*, pp. 1-44, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Sébastien Bubeck et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4 (Version 1)," *arXiv:2303.12712*, pp. 1-155, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Zahra Atf et al., "The Challenge of Uncertainty Quantification of Large Language Models in Medicine (Version 1)," *arXiv:2504.05278*, pp. 1-25, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jason Wei et al., "Emergent Abilities of Large Language Models," *Transactions on Machine Learning Research*, pp. 1-30, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Vibhor Agarwal et al., "MedHalu: Hallucinations in Responses to Healthcare Queries by Large Language Models (Version 2)," *arXiv:2409.19492*, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Josh Achiam et al., "GPT-4 Technical Report (Version 1)," *arXiv:2303.08774*, pp. 1-100, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yubo Ma et al., "SciAgent: Tool-Augmented Language Models for Scientific Reasoning," *arXiv:2402.11451*, pp. 1-34, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yujia Qin et al., "Tool Learning with Foundation Models (Version 1)," *arXiv:2304.08354*, pp. 1-75, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Sijia Chen et al., "CARES: Comprehensive Evaluation of Safety and Adversarial Robustness in Medical LLMs (Version 1)," *arXiv:2505.11413*, pp. 1-31, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] Harsha Nori et al., "Capabilities of GPT-4 on Medical Challenge Problems (Version 1)," *arXiv:2303.13375*, pp. 1-35, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Guangyu Wang et al., "ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation (Version 1)," *arXiv:2306.09968*, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730-27744, 2022. [Google Scholar] [Publisher Link]

[15] Yexiao He et al., "MedOrch: Medical Diagnosis with Tool-Augmented Reasoning Agents for Flexible Extensibility (Version 1)," *arXiv:2506.00235*, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Congzhen Shi et al., "Towards Trustworthy Foundation Models for Medical Image Analysis," *arXiv:2407.15851*, pp. 1-44, 2024. [Google Scholar] [Publisher Link]