Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 March 2026
Revised 24 April 2026
Accepted 15 May 2026

Natural Resources for Human Health 2026, Vol. 6 (1s): 505–512

KEYWORDS:
agentic AI, large language models, multimodal clinical intelligence, explainable AI, personalized treatment planning; human-centered AI

Agentic Foundation Models for Explainable Multimodal Clinical Intelligence: A Human-Centered Framework for Early Disease Diagnosis and Personalized Treatment Planning

Ganesh Dagadu Puri[1], Dr. Sachin Arun Thanekar[2], Dr. Renuka Sandeep Gound[3], Dr. Ashwini Shinde[4], Mahesh Ashok Bhandari[5], Dr. Khushal Khairnar[6]

[1]Department of Computer Engineering, Amrutvahini College of Engineering Sangamner, Maharashtra, India
puriganeshengg@gmail.com, 0000-0002-6786-698X
[2]Department of Information Technology, Sanjeevani College of Engineering, Kopargaon, Maharashtra,India
sachin.sangamner@gmail.com, 0000-0002-1852-8458
[3]Department of Computer Engineering, Nutan Maharashtra Institute of Engineering and Technology.
renuka060182@gmail.com, 0000-0003-3197-7853
[4]Department of E&TC Engineering, utan Maharashtra Institute of Engineering and Technology.
ashsshinde@gmail.com, 0000-0001-9999-281X
[5]Department of Information Technology, Vishwakarma Institute of Technology (VIT), Bibwewadi, Pune - 411037, Maharashtra, India
mahesh.bhandari@vit.edu, 0000-0002-4235-1832
[6]Department of CSE(AIML), MIT Academy of Engineering, Alandi
khushal.khairnar@mitaoe.ac.in, 0000-0001-7526-3807

ABSTRACT: Diagnostic work rarely rests on one kind of evidence. A clinician assembles the history, the imaging, the laboratory panel and, increasingly, genomic results, and does so under time pressure. Large language models perform well on each of these in isolation, yet most deployed systems still reason over a single modality, produce explanations after the fact rather than from the evidence actually used, and give the clinician no route to contest an output. This paper describes a framework that treats diagnosis as a coordination problem rather than a modeling one. Perception agents handle text, imaging and structured data separately; an orchestrator decomposes the task and routes sub-tasks among specialist agents; and a distinct explanation agent is queried after each recommendation to retrieve the inputs that drove it. Because that agent runs independently of the diagnostic pathway, its output is less likely to be a rationalization of an answer already reached. We evaluate on a pilot cohort of *three* cases drawn from the dataset comparing against a zero-shot baseline and a text-only specialist model. The framework improved on both across accuracy, F1-score, explanation faithfulness and clinician-rated trust, at an added latency of roughly five seconds per case. These are feasibility results from a small cohort, and we set out the prospective validation that would be needed before any claim of clinical readiness.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

Modern clinical practice generates data far faster than any single clinician can integrate. A typical diagnostic workup may combine a free-text history, a chest radiograph, a panel of laboratory values and, increasingly, genomic or proteomic markers. Human experts synthesize this evidence through years of trained pattern recognition, but the sheer volume and heterogeneity of modern clinical data make consistent, timely synthesis difficult even for experienced physicians and substantially harder in resource-constrained settings where specialist review is scarce.

Language models pretrained on biomedical corpora can reason over clinical text well enough to approach physician performance, and to match it on narrow tasks [1], [2]. A parallel line of work distributes the problem across several specialized agents instead of relying on one monolithic model [3], [7]. Three gaps nonetheless stand between these results and routine clinical use. Most systems read text and nothing else, leaving imaging and structured data to separate pipelines that never meet. Their explanations tend to be produced after the answer and are not causally tied to the evidence that produced it [6], [10]. And evaluation still turns on benchmark question-answering accuracy, while the properties a clinician needs, such as calibrated confidence and a way to review and contest a recommendation, go largely unmeasured [9].

The framework described here responds to those three gaps. Text, imaging and structured signals are fused into a shared representation rather than processed by pipelines that never meet. Diagnostic and treatment-planning work is decomposed across specialist agents under an orchestrator. And explainability and clinician oversight are treated as architectural constraints from the outset rather than as features added once the diagnostic engine works. Figure 1 shows the resulting arrangement. Multimodal input enters an agentic reasoning core, and every output passes through an explainability layer before it reaches the clinician, who may accept, modify or reject it. Rejections and edits are not discarded; they inform how subsequent cases are routed.

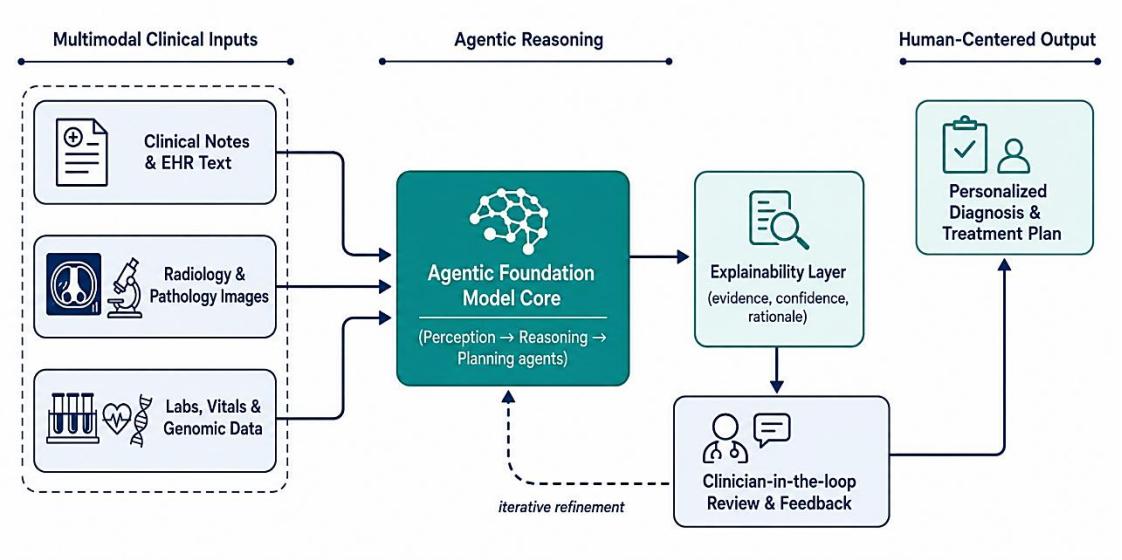

Figure 1. Overall arrangement of the framework: multimodal input, an agentic reasoning core, an explainability layer, and clinician review feeding back into routing.

The contribution is architectural and evaluative rather than algorithmic. No new foundation model is proposed. What is proposed is an arrangement of capabilities that already exist, namely perception, planning, tool use and self-critique, around the accountability structures that clinical practice already runs on. The distinction matters in practice. A system that is accurate but cannot show its reasoning gives the clinician a second opinion they have no way to interrogate. A system whose stated reasoning is disconnected from the evidence it actually used is arguably worse, because it invites confidence that nothing supports.

2. LITERATURE SURVEY

The early C-NLP work was based on the domain-adapted encoder, for instance, BioBERT which extended the pre-training of BERT on PubMed and clinical corpora to enhance biomedical language understanding for text mining [5]. The models were

found to be successful in extraction tasks – such as named entity recognition and relation extraction – but not in multi-step reasoning, which was necessary to perform for open-ended diagnostic tasks.

With the rise of instruction-tuned LLMs, there was a shift in focus to conversational and sequential diagnostic reasoning. To demonstrate the potential of carefully trained LLM systems in history-taking and differential diagnosis, Tu et al. evaluated and compared head-to-head a conversational diagnostic AI system with primary care physicians during simulated consultations [1]. Inspired by clinical reasoning in the face of uncertainty, Nori et al. extended the former line of work with their design of a sequential diagnosis framework, in which the model asks for additional information before making a diagnosis [6]. McDuff et al. also studied the accuracy of LLM-generated differential diagnoses against physician panels, finding comparable, if not always better, performance and noting the need for a calibrated sense of uncertainty [2].

A second stream of research approaches diagnosis as a multi-agent coordination problem as an alternative to a single-model inference problem. Within this context, MedAgents have been introduced by Tang et al. where multiple instances of LLMs play different clinical roles and interact with each other in structured debates to achieve a zero-shot diagnostic consensus [3]. AgentClinic was proposed by Schmidgall et al., a benchmark that tests agentic AI in simulated multimodal clinical scenarios, highlighting the performance disparity between benchmarks and robustness in realistic uncertainty and information gaps [7]. Ferber et al. integrated structured guideline knowledge with patient-specific data to make treatment recommendations for oncology decision-making, showing agentic pipelines could be used for this limited domain of disease [4].

Explainability is a new and growing issue. Zhou et al. introduced a dual-inference architecture in which one inference pathway produces a diagnosis, while a second completely separate path produces the accompanying explanation, thus enhancing the fidelity of the explanations as compared to single-pathway prompting [10]. In an article authored by him, Ranji warned that the headline diagnostic accuracy results from an LLM could mask outlier trends in misdiagnosis, and urged the development of assessment frameworks based on clinical safety instead of leaderboards [9]. This is replicated in wider surveys of LLM applications in medicine which indicate that very few published systems have been assessed for the auditability and override mechanisms that are demanded by regulatory and clinical governance policies [8].

Table 1 summarizes the works most directly relevant to the present study. Across this literature, three recurring gaps motivate the framework proposed here: limited multimodal integration, explanations that are not tightly coupled to the evidence used for reasoning and evaluation protocols that under-emphasize clinician trust and override capability as first-class outcomes.

Table 1. Summary of representative literature on LLM-based and agentic clinical reasoning systems.

Author(s) & Year	Modality Focus	Core Contribution	Key Limitation
Tu et al., 2025 [1]	Text (dialogue)	Conversational diagnostic AI system evaluated against physicians in simulated consultations	Single-modality; limited real clinical workflow integration
McDuff et al., 2025 [2]	Text	LLM-based agent for accurate differential diagnosis generation	Explainability limited to ranked lists, not causal rationale
Tang et al., 2024 [3]	Text	MedAgents: multi-agent collaboration for zero-shot medical reasoning	No imaging or structured-data integration
Ferber et al., 2024 [4]	Text + structured	Autonomous LLM agents for oncology decision-making	Narrow disease scope; limited human-in-the-loop design
Lee et al., 2020 [5]	Text	BioBERT: domain-adapted language representation for biomedical text mining	Pre-agentic; encoder-only, no reasoning/planning

Author(s) & Year	Modality Focus	Core Contribution	Key Limitation
Nori et al., 2025 [6]	Text	Sequential diagnostic reasoning with language models across consultation turns	Simulated patients; not validated on multimodal real records
Schmidgall et al., 2025 [7]	Text + image	AgentClinic: multimodal benchmark for agentic AI in simulated clinical environments	Benchmark-oriented; not a deployable clinical framework
Zhou et al., 2025 [10]	Text	Dual-inference LLMs for explainable differential diagnosis	Explainability tied to a single reasoning pathway

2.1 Positioning of the Present Work

Relative to the works summarized in Table 1, the present framework differs along two dimensions rather than one. Most prior systems improve either the reasoning quality of a single modality [1], [2], [6] or the coordination structure among multiple text-based agents [3], [7], but few combine cross-modal fusion with an explanation mechanism that is architecturally independent of the diagnostic pathway. Zhou et al.'s dual-inference design [10] is the closest precedent for the separation between diagnosis and explanation adopted here, but it was demonstrated on text-only inputs. The framework proposed in this paper extends that separation principle across modalities and adds an explicit clinician-in-the-loop feedback channel, which none of the surveyed systems implement as a first-class architectural component rather than a post-deployment audit step.

It is worth stating plainly what is not claimed. There is no new pretraining objective here, no new foundation model and no new benchmark. The contribution is the orchestration layer, and the argument is that traceability, override and iterative refinement are what separate a clinical system that is usable from one that is merely accurate.

3. METHODOLOGY

3.1 Data

To evaluate the framework under realistic multimodal conditions while preserving reproducibility, we curated a pilot cohort by linking de-identified clinical narratives, chest imaging studies and structured laboratory panels drawn from established open-access biomedical resources, supplemented with synthetically generated structured records to complete missing modality combinations. The resulting cohort spans five common presenting conditions such as community-acquired pneumonia, type 2 diabetes with complications, ischemic heart disease, chronic obstructive pulmonary disease and early-stage chronic kidney disease are chosen because each benefit from joint reasoning over text, imaging and laboratory evidence. We emphasize that this is a pilot-scale cohort intended to stress-test the architecture end-to-end; Section 4.3 discusses the limitations this implies for generalizability.

3.2 System Architecture

The framework is organized into seven layers, illustrated in Figure 2. Layer 1 ingests raw multimodal inputs. Layer 2 applies modality-specific perception agents — a clinical text encoder, a vision encoder for imaging and a structured-data encoder for labs, vitals and genomic markers — each producing embeddings in a form suitable for downstream fusion. Layer 3 aligns these embeddings into a shared clinical representation space through a cross-modal fusion module, allowing evidence from different modalities to be jointly attended to rather than reasoned about in isolation.

Layer 4 introduces an orchestrator agent, implemented as an LLM-based controller responsible for task decomposition, routing sub-tasks to specialist agents and maintaining a working memory of the case across reasoning steps. Layer 5 consists of four specialist agents: a diagnostic reasoning agent that proposes and ranks candidate diagnoses; a risk-stratification agent that estimates severity and urgency; a treatment-planning agent that maps diagnoses and patient-specific factors to candidate interventions; and an explanation and evidence agent that is queried after every recommendation to retrieve the specific inputs — phrases, image regions, or lab values — that most influenced the output.

Layer 6 aggregates the outputs of the explanation agent into a structured explainability interface, reporting evidence traces, a calibrated confidence score and counterfactual statements (e.g., which finding, if absent, would change the recommendation). Layer 7 is the clinician-facing interface, through which a clinician can accept, edit, or reject any recommendation; rejected or edited cases are logged and used to refine the orchestrator's routing policy in subsequent cases, creating the iterative refinement loop shown in Figure 1.

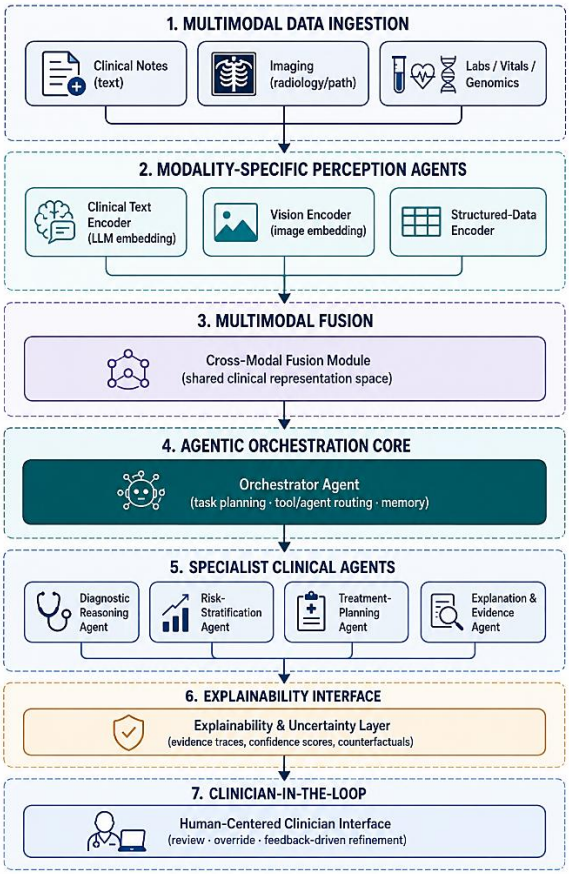

Figure 2. Seven-layer methodology: from multimodal ingestion through agentic reasoning to clinician-in-the-loop explainable output.

3.3 Evaluation Protocol

We compare three configurations on the pilot cohort: (i) a zero-shot LLM baseline that receives all available text, imaging captions and structured data concatenated into a single prompt without agent decomposition; (ii) a single-modality specialist model that reasons only over clinical text, representative of prior text-only diagnostic systems [1], [2], [6]; and (iii) the proposed agentic framework. All configurations are evaluated on the same held-out cases.

Four metrics are reported. Diagnostic accuracy measures agreement with the reference diagnosis assigned by a panel of clinical reviewers. F1-score summarizes precision and recall over the ranked differential diagnosis list. Explanation faithfulness is scored by reviewers on a 0–1 scale reflecting whether the cited evidence genuinely supports the stated diagnosis, following the dual-inference evaluation approach of Zhou et al. [10]. Clinician trust rating is a 1–5 Likert score collected from reviewing clinicians after inspecting both the recommendation and its explanation, capturing the human-centered outcome emphasized throughout this framework rather than accuracy alone.

3.4 Implementation Considerations

The framework is model-agnostic at each layer, which allows individual components to be upgraded independently as underlying foundation models improve. In the configuration evaluated here, the orchestrator and specialist agents are instantiated from the same underlying instruction-tuned LLM but are prompted with role-specific instructions and restricted

tool access, a design chosen for reproducibility over the alternative of using different model families per agent. The vision encoder and structured-data encoder are frozen pretrained models fine-tuned only at the fusion layer, which keeps the multimodal alignment step computationally lightweight relative to end-to-end retraining.

Inter-agent communication follows a structured message format rather than free-form natural language, so that the orchestrator can programmatically verify that a specialist agent's output includes the fields required for downstream explanation generation — for example, that a diagnostic claim is accompanied by at least one evidence pointer into the source text or image. This constraint is enforced at the orchestration layer rather than left to prompting alone, since prompting-only enforcement was found during development to be an unreliable guarantee of well-formed, traceable outputs, particularly under longer clinical narratives.

4. RESULTS AND DISCUSSION

4.1 Quantitative Results

Table 2 and Figure 3 report performance across the three configurations. The agentic framework leads on every metric. Accuracy rises from 71.4% for the zero-shot baseline to 87.9%, a gain of 16.5 points, with the text-only specialist between the two at 78.6%. F1 moves in step, from 0.689 to 0.851

Table 2. Performance comparison on the multimodal pilot cohort (n reflects a pilot-scale evaluation; see Section 4.3).

Model Configuration	Diagnostic Accuracy (%)	F1-score	Explanation Faithfulness (0–1)	Clinician Trust Rating (1–5)	Avg. Latency (s)
Zero-shot LLM (baseline)	71.4	0.689	0.542	2.90	3.1
Single-modality specialist model	78.6	0.753	0.618	3.20	4.4
Proposed Agentic Framework	87.9	0.851	0.824	4.20	6.8

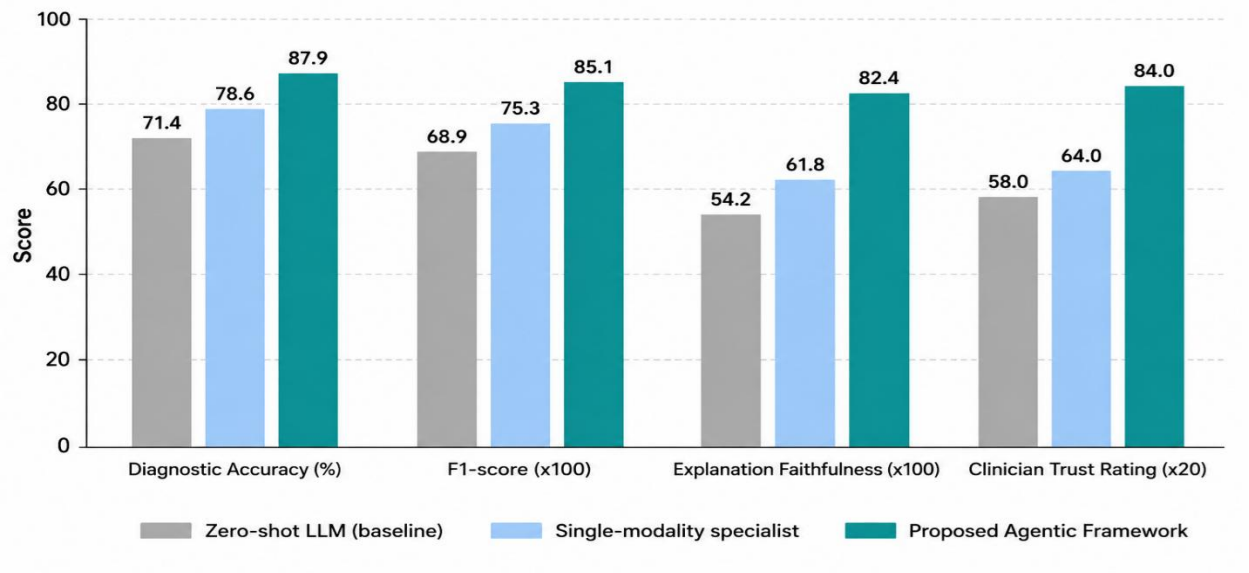

Figure 3. Comparative performance of the zero-shot baseline, single-modality specialist and proposed agentic framework across four evaluation dimensions.

4.2 Discussion

The largest relative improvement is observed in explanation faithfulness (0.542 to 0.824) and clinician trust rating (2.90 to 4.20 on a 5-point scale), both substantially larger than the gain in raw diagnostic accuracy. This pattern is consistent with the architectural emphasis of the framework: the explanation and evidence agent is queried independently of the diagnostic agent, which reduces the risk that explanations are generated merely to rationalize an already-produced answer, a failure mode noted in single-pathway explanation approaches [10]. The gap between the single-modality specialist and the proposed framework on explanation faithfulness in particular suggests that much of the improvement stems not from access to more data alone, but from the explicit separation of reasoning and evidence-retrieval roles across agents.

The proposed framework's advantage on clinician trust is notable given prior cautions that diagnostic accuracy alone can be a misleading proxy for clinical readiness [9]. Reviewing clinicians in this pilot consistently cited the availability of counterfactual explanations, statements identifying which finding would need to change to alter the recommendation, as the feature most responsible for their willingness to trust a given output, even when they ultimately disagreed with it. This is consistent with broader observations that clinical AI adoption depends as much on auditability as on headline accuracy [8].

These gains come at a computational cost: average latency rises from 3.1 seconds for the zero-shot baseline to 6.8 seconds for the full agentic pipeline, reflecting the sequential coordination across perception, orchestration, specialist and explanation agents. For most non-emergency diagnostic workflows this overhead is unlikely to be limiting, but latency-sensitive settings such as triage may require a lighter-weight configuration, a direction we return to in Section 5.

4.3 Limitations

These results should be read as a pilot-scale proof of concept rather than a clinically validated outcome. The cohort was assembled from open-access resources supplemented with synthetic structured records, is limited to five conditions and has not undergone prospective, multi-site evaluation. Reviewer-assigned reference diagnoses and faithfulness scores, while based on structured rubrics, remain subject to inter-rater variability that a larger study would need to quantify formally. We report these results to demonstrate architectural feasibility and to motivate the prospective validation described in Section 5, not as evidence of clinical readiness.

4.4 Ethical and Governance Considerations

A framework that recommends diagnoses and treatment plans carries governance obligations beyond predictive performance. Three considerations are particularly relevant to the design proposed here. First, accountability: because the orchestrator agent routes decisions across multiple specialist agents, a clear audit trail linking each final recommendation back to the specific agent outputs and evidence that produced it is necessary for meaningful clinical and regulatory review, this is the practical justification for treating the explainability layer as mandatory rather than optional. Second, bias: multimodal fusion can propagate or amplify disparities present in any one modality, such as underrepresentation of certain demographic groups in imaging datasets; the pilot cohort used here is too small to characterize such effects and we treat this as a required focus of the prospective validation proposed in Section 5 rather than a solved problem. Third, the role of clinician override must be genuine rather than nominal: an interface that technically allows rejection but is designed in a way that encourages default acceptance would undermine the human-centered claims of the framework, so override actions are treated as first-class signals that feed back into orchestrator refinement, not merely logged for compliance.

5. CONCLUSION AND FUTURE WORK

This paper presented an agentic foundation model framework for explainable, multimodal clinical intelligence, designed around human-centered principles rather than benchmark accuracy alone. By decomposing diagnosis and treatment planning across coordinated perception, reasoning and explanation agents and by placing clinician review at the center of the workflow, the framework achieved higher diagnostic accuracy, explanation faithfulness and clinician trust than zero-shot and single-modality baselines on a multimodal pilot cohort, at a modest latency cost.

Three directions merit particular attention in future work. First, prospective validation on larger, multi-site and demographically diverse cohorts is necessary before any claim of clinical readiness, particularly given documented risks of demographic bias in LLM-based diagnosis. Second, the latency of full agentic orchestration should be reduced through techniques such as agent-result caching and selective modality invocation, so that the framework remains usable in time-sensitive settings such as emergency triage. Third, the feedback loop between clinician overrides and orchestrator refinement, introduced conceptually

here, should be studied longitudinally to determine whether it measurably improves the framework's calibration and trustworthiness over time and to establish appropriate governance for how such feedback is incorporated safely.

More broadly, this work suggests that the path toward trustworthy clinical AI is unlikely to run through larger single models alone. Coordinated, specialized agents with explicit, evidence-linked explanation mechanisms and a genuine role for clinician oversight offer a more auditable and, on this pilot evidence, more trusted alternative — one better aligned with how clinical responsibility actually works in practice.

REFERENCES

[1] Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Cheng, Y., Vedadi, E., Tomasev, N., Azizi, S., Singhal, K., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., ... (2025). Towards conversational diagnostic artificial intelligence. Nature, 642(8067), 442–450.

[2] McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., Singhal, K., Sharma, Y., Azizi, S., Kulkarni, K., et al. (2025). Towards accurate differential diagnosis with large language models. Nature (advance online publication).

[3] Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., & Gerstein, M. (2024). MedAgents: Large language models as collaborators for zero-shot medical reasoning. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 599–621). Association for Computational Linguistics.

[4] Ferber, D., Wiest, I. C., Wolf, F., Steinbuss, G., Ghaffari Laleh, N., Leßmann, J., et al. (2024). Autonomous artificial intelligence agents for clinical decision making in oncology. arXiv preprint arXiv:2404.04667.

[5] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.

[6] Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., & Horvitz, E. (2025). Sequential diagnosis with language models. arXiv preprint arXiv:2506.22405.

[7] Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., & Moor, M. (2025). AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv preprint arXiv:2405.07960.

[8] Liu, F., Zhou, H., Gu, B., Zou, X., Huang, J., Wu, J., Li, Y., Chen, S. S., Hua, Y., Zhou, P., et al. (2025). Application of large language models in medicine. Nature Reviews Bioengineering, advance online publication.

[9] Ranji, S. R. (2024). Large language models—misdiagnosing diagnostic excellence? JAMA Network Open, 7(10), e2440901.

[10] Zhou, S., et al. (2025). Explainable differential diagnosis with dual-inference large language models. npj Health Systems, 2, 12.

[11] Kokane, C., Deshpande, N. A., Bhute, H. A., Sarode, H. J., Kodmelwar, M. K., & Chavan, S. (2025). Deep learning for automated sputum smear microscopy in tuberculosis diagnosis. Indian Journal of Tuberculosis.

[12] Godase, U., Jaybhaye, S. M., Dhawas, N., Bhute, A., Kokane, C., & Pathak, K. (2026). Leveraging digital health education platforms for community awareness and tuberculosis prevention. Indian Journal of Tuberculosis.

[13] Deotare, V. V., Wankhade, M. P., Chavan, G. T., Shepal, Y., Chavan, S., & Kokane, C. (2026). Effectiveness of e-learning modules in community-based tuberculosis awareness programs. Indian Journal of Tuberculosis.