

From Static Symptom Checking To Adaptive Dialogue: Design And Evaluation Of An Agentic Llm System For First-Level Medical Triage

Venkata Reddy Konasani*¹, Dr. M.R. Jhansi Rani²

¹Research Scholar, FPM, ISBR Business School

Email ID: 21.venkat@gmail.com

²Professor & Director, Research Centre, ISBR Business School

Email ID: dr.jhansirani@isbr.in

***Corresponding author**

Venkata Reddy Konasani

Research Scholar, FPM, ISBR Business School

Email ID: 21.venkat@gmail.com

Cite this paper as: Venkata Reddy Konasani, Dr. M.R. Jhansi Rani, (2026) Consumer Wages and Remittance Behaviour among Migrant and Local Labour in Kerala. *Frontiers in Health Informatics*, Vol.15, No 1.2026, 19-30

ABSTRACT

First-contact medical assessment requires more than generation of a plausible diagnostic label. Safe support depends on gathering sufficient information, distinguishing low-risk from high-risk presentations, recognising red flags and directing the user toward an appropriate level of care. Conventional symptom checkers commonly use fixed pathways, while large language models can interpret free-text descriptions but may reach conclusions before enough information is available. This study presents the design and simulation-based comparative evaluation of LLMDoctor, an agentic conversational prototype for first-level medical assessment and triage support. The architecture combines natural-language symptom entry, information-gap detection, adaptive follow-up questioning, progressive context updating, preliminary differential assessment, red-flag screening, triage recommendation and human escalation. A comparative case-based evaluation framework contrasts a conventional symptom checker, a single-turn LLM baseline and the agentic condition. A structured simulation comprising 120 constructed cases was used to demonstrate the comparative evaluation framework across diagnostic performance, triage agreement, under-triage, red-flag recognition, interaction quality and information completeness. The article contributes a bounded agentic decision-support architecture, separates diagnostic performance from triage safety and proposes a safety-error taxonomy for evaluating first-level medical AI. The framework positions escalation as a core system capability and retains clinician authority for high-risk or uncertain cases.

Keywords: agentic artificial intelligence, large language models, medical triage, adaptive dialogue, symptom assessment, red-flag detection, clinical decision support, human oversight.

INTRODUCTION

First-level medical assessment begins before a definitive diagnosis is available. A person may present with a short, imprecise description such as persistent tiredness, abdominal discomfort, dizziness or breathlessness, and the first task is to determine what information is missing and whether the presentation may require urgent care. The practical decision is not limited to naming a disease. It includes identifying relevant history, recognising warning signs, deciding whether further questioning is necessary and recommending an appropriate next step. Digital systems that operate at this point of contact can improve access to preliminary information, yet mistakes in urgency classification can create direct safety consequences. Under-triage may delay needed treatment, while excessive over-triage can increase unnecessary use of emergency services. For this reason, first-contact medical AI needs to be evaluated as a decision-support process rather than as a question-answering tool alone (Semigran et al., 2015; Wallace et al., 2022). Digital symptom checkers were an early attempt to support self-assessment at scale. These systems typically collect symptoms and demographic information, apply fixed decision rules or probabilistic logic and return a list of possible conditions together with care advice. Their accessibility made them attractive for

patient-facing use, yet large evaluations reported wide variation in both diagnostic and triage performance. Semigran et al. (2015) found limited primary diagnostic accuracy across publicly available tools, while Chambers et al. (2019) and Wallace et al. (2022) documented persistent inconsistency across platforms and clinical presentations. The fixed-pathway structure also creates an interaction constraint. A predetermined sequence can be efficient for familiar presentations, but it may not adapt well when patient narratives are vague, incomplete or contain several possible concerns. The clinical value of the interaction depends on whether the system can recognise that it lacks enough information before issuing a recommendation. Large language models introduced a different interaction model because users can describe symptoms in ordinary language and receive context-sensitive responses. Early demonstrations of medical knowledge were striking. ChatGPT reached passing-level performance on United States Medical Licensing Examination questions in the study by Kung et al. (2023), and medically aligned systems such as Med-PaLM 2 achieved much stronger benchmark performance with physician-rated improvements in long-form medical question answering (Singhal et al., 2025). These results established that LLMs can encode substantial clinical knowledge and generate coherent medical explanations. They did not establish readiness for autonomous clinical decision-making. Medical examinations provide the relevant information inside a structured question, whereas real first-contact assessment often begins with incomplete information that must be actively gathered. Evidence from more realistic evaluations demonstrates the importance of this distinction. Hager et al. (2024) evaluated LLMs across a clinical decision pathway using real-world cases and found reduced performance when models had to gather information rather than receive the full case directly. The study also identified failures in guideline adherence, laboratory interpretation, instruction following and robustness to the amount and order of clinical information. In emergency-department comparisons, Fraser et al. (2023) observed a similar separation between diagnostic and triage capability. ChatGPT could produce plausible differential diagnoses, yet unsafe urgency recommendations remained a major concern. The result supports a central premise for first-level AI evaluation: diagnostic plausibility and triage safety are separate properties, and strong performance on one cannot be treated as evidence of the other. Agentic AI provides a pathway for addressing the information-gathering problem. Instead of treating a patient statement as a complete case, an agentic system can maintain a working representation of the conversation, identify missing information, select a follow-up question and update its assessment after each response. The conversational diagnostic AI work reported by McDuff et al. (2025) demonstrates the value of multi-turn clinical dialogue and evaluates history-taking, diagnostic accuracy, management, communication and empathy as distinct dimensions. Related work on differential diagnosis shows that interactive LLM interfaces can support iterative clinical reasoning rather than isolated answer generation (Tu et al., 2025). More recent full-process agentic research extends this idea by allowing the system to begin with a vague chief complaint, seek further information and request human assistance for clinical operations that cannot be performed by software (Xu et al., 2026). The shift toward agentic medical AI creates a new design problem. Greater autonomy can improve information acquisition, yet autonomy also increases the number of decisions that can propagate an early mistake. A poorly chosen follow-up question may omit a critical symptom. An overconfident provisional diagnosis may narrow the later search. A triage recommendation may become unsafe even when the differential contains the correct condition. The relevant goal is not unrestricted autonomy. It is bounded delegation, where the agent performs defined information-processing tasks while the system retains explicit conditions for escalation. In a high-stakes setting, safe design requires the agent to recognise uncertainty, preserve alternative explanations, prioritise red flags and transfer decision authority when clinical examination, testing or professional judgement is required. LLMDoctor is a conversational research prototype developed to examine bounded agentic support for first-level medical assessment. The documented demonstration shows a user entering a natural-language concern, the bot asking follow-up questions, and the later response presenting a summary, possible causes, general self-care suggestions and red-flag information. The project objectives include an LLM-powered dialogue agent, integration with structured medical ontologies and symptom databases, safe triage, diagnostic evaluation and assessment of patient interaction quality. The current prototype documentation does not specify the ontology, database, model provider or retrieval architecture. These components are consequently represented at the functional level rather than attributed to specific technologies. The present article keeps those boundaries explicit and describes only the documented interaction concept, together with a rigorous evaluation framework that can be applied once the technical implementation is frozen. The research gap is fourfold. First, many patient-facing systems still evaluate a symptom list rather than the quality of adaptive information acquisition. Second, studies often report diagnostic performance without giving equal analytical weight to under-triage, red-flag misses and escalation decisions. Third, agentic dialogue introduces safety questions that are not captured by Top-k diagnostic accuracy alone, including whether the agent asks necessary questions, avoids premature closure and stops at an appropriate point. Fourth, human oversight is frequently described as a deployment recommendation rather than represented as an explicit system function with observable escalation criteria. These gaps call for an evaluation approach that treats dialogue, diagnosis, triage and safety as connected but non-interchangeable layers. The study aims to design and evaluate an agentic LLM-based conversational system for first-level medical assessment and triage support, with attention to adaptive information gathering, preliminary differential reasoning, red-flag recognition, triage appropriateness,

communication quality and human escalation. The contribution is not a claim of autonomous medical diagnosis. It is a design and evaluation framework for determining whether adaptive dialogue adds measurable value over static interaction while preserving safety boundaries. The article contributes an agentic assessment architecture, a comparative case-based protocol, a multidimensional outcome framework and a safety-error taxonomy that can support transparent evaluation of future patient-facing agentic systems.

BACKGROUND AND RELATED WORK

From Rule-Based Symptom Checking to Conversational Assessment

Computer-assisted diagnosis has a long history in clinical decision support, but patient-facing symptom assessment became widely visible through web and mobile symptom checkers. These systems translate user-selected symptoms into possible conditions and urgency advice. The approach offers consistency and scalability, yet its performance depends heavily on the coverage of the underlying rules, disease models and branching logic. Semigran et al. (2015) reported that the correct diagnosis was listed first in only a minority of standardised cases, while later systematic reviews found broad variation in primary diagnosis and triage accuracy (Chambers et al., 2019; Wallace et al., 2022). These findings are important for system design because low diagnostic accuracy and inaccurate triage create different consequences. A missed diagnosis may reduce informational value, while unsafe under-triage can delay urgent care. The static nature of conventional symptom checking is also relevant. A decision tree can ask follow-up questions, but the questions are drawn from a predefined path. This design is efficient when patient information maps cleanly to the decision structure. It becomes less flexible when users provide non-standard language, several symptoms at once or partial information. General practice relies heavily on history-taking because clinicians use each response to decide what question should come next. An adaptive dialogue system aims to approximate that information-acquisition process without assuming that the first message contains a complete diagnostic picture.

Large Language Models in Medical Reasoning

LLMs changed medical natural-language processing by allowing the same model to interpret free-form narratives, answer medical questions, summarise information and generate differential diagnoses. The USMLE evaluation by Kung et al. (2023) showed that a general-purpose conversational model could reach or exceed passing-level performance across several examination formats. Singhal et al. (2023) showed strong encoded clinical knowledge in a medically aligned LLM, while the later Med-PaLM 2 study reported higher benchmark accuracy and improved physician-rated response quality across several clinical dimensions (Singhal et al., 2025). These studies established a high ceiling for medical question answering, yet benchmark questions provide a more complete information package than real patient encounters.

The difference becomes visible when a model must perform the intermediate work of clinical decision-making. Hager et al. (2024) showed that leading LLMs struggled when required to gather information, follow clinical guidelines, interpret laboratory results and maintain robustness across changes in information quantity and order. Goh et al. (2024) also found that access to an LLM did not automatically improve clinician diagnostic reasoning in a randomized trial, even though the model demonstrated substantial standalone capability. These findings indicate that raw language-model competence and effective human-AI clinical workflow are separate concerns.

Agentic AI and Adaptive Clinical Dialogue

Agentic systems extend LLM use from response generation to sequential task execution. In a medical dialogue, an agent can identify uncertainty, decide what information to request next, use external resources, update a working hypothesis and determine whether the case should be escalated. McDuff et al. (2025) developed AMIE as a system optimised for diagnostic dialogue and evaluated it through multi-turn encounters with validated patient actors. The evaluation separated history-taking, diagnostic reasoning, management and communication, demonstrating why a medical conversation cannot be reduced to one accuracy score. The work also showed that diagnostic dialogue can be systematically assessed using both clinician-centred and patient-centred criteria.

Agentic reasoning is moving closer to full-process clinical support. Gaber et al. (2025) evaluated LLM workflows for triage, specialty referral and diagnosis using a large curated clinical dataset and reported that richer clinical context and a retrieval-assisted workflow could improve selected tasks. Xu et al. (2026) presented a full-process diagnosis architecture that begins with a vague complaint and progressively acquires information, while requesting physician assistance for examinations or tests that software cannot perform. The architecture is especially relevant to bounded autonomy because it preserves human intervention at points where clinical operations or risk management are required. These studies support a shift from asking whether an LLM knows medicine toward asking whether an agent can manage the sequence of information acquisition and escalation safely.

Triage Safety, Red Flags and Human Oversight

Triage is a distinct safety task because the system must translate an uncertain presentation into an action level.

Fraser et al. (2023) reported meaningful differences between diagnostic and triage performance across ChatGPT, symptom checkers and physicians. Wallace et al. (2022) argued that triage may be more important than exact diagnosis for patient-facing symptom tools because the immediate objective is often to guide the user toward an appropriate care pathway. This distinction motivates separate outcome measures for exact triage agreement, under-triage and over-triage. Under-triage deserves particular attention because it may falsely reassure a user with a serious condition. Human oversight should be represented as part of the system architecture rather than as a generic statement at the end of an article. Hager et al. (2024) found that current models may require extensive supervision when they fail to follow instructions or request appropriate examinations. Xu et al. (2026) placed clinician assistance inside the diagnostic process when physical examination, laboratory testing or other clinical operations were needed. This design logic supports a bounded agentic model in which software can collect, organise and reason over information, while clinicians retain authority for high-risk interpretation, testing and final decisions.

RESEARCH GAP AND DESIGN REQUIREMENTS

The literature indicates that the next research problem is not simply whether an LLM can produce a plausible diagnosis. The harder question is whether an agent can manage the information-gathering and triage process without creating new safety risks. The present design translates this problem into seven requirements that can be evaluated directly. Adaptive information acquisition: The system should recognise missing information and choose follow-up questions that are relevant to the presenting concern rather than rely only on a fixed checklist.

Progressive case updating: Each user response should modify the working case representation so later reasoning is based on the conversation accumulated to that point. Differential rather than definitive output: The system should preserve plausible alternatives and avoid presenting a preliminary inference as a confirmed diagnosis.

Independent triage assessment: Urgency should be evaluated as a separate decision layer because diagnostic plausibility does not guarantee safe care-level recommendation. Red-flag priority: Warning signs should receive explicit priority so that a potentially dangerous presentation can trigger escalation even when diagnostic confidence is incomplete. Uncertainty and scope control: The system should communicate limits, recognise when evidence is insufficient and avoid unsupported precision. Human escalation: The workflow should include explicit transfer conditions when risk, uncertainty, required examination or treatment decisions exceed the appropriate scope of a patient-facing system. These requirements provide the bridge between prior medical AI research and the LLMDoctor architecture. They also create measurable evaluation criteria. A system that asks more questions is not necessarily better. The relevant question is whether those questions improve information completeness, reduce unsafe triage and support a safer decision pathway.

System Architecture and Agentic Workflow

LLMDoctor is treated in this article as a first-level medical assessment and triage-support prototype. The available demonstration confirms natural-language user input, multi-turn follow-up questioning and a structured patient-facing response. The system design includes planned integration with structured medical ontologies and symptom databases. As the implemented resources are not yet fixed, the architecture represents these functions through a technology-neutral medical knowledge-support layer.

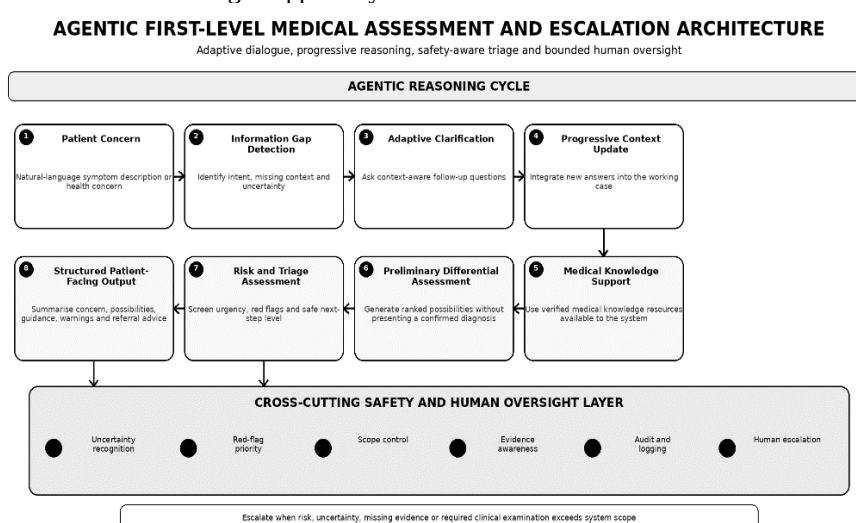


Figure 1. Agentic first-level medical assessment and escalation architecture.

Source: Developed by the author

Patient Concern

The interaction begins with a natural-language description rather than a complete structured case. This preserves the patient-facing character of the tool and allows users to describe symptoms in their own words. The first message is treated as an initial signal, not as sufficient evidence for a diagnosis.

Information-Gap Detection

The agent interprets the complaint and identifies what clinically relevant context is missing. The objective is to prevent premature closure. Missing duration, severity, associated symptoms, medication context or warning signs may change the next question and the urgency assessment.

Adaptive Clarification

The agent selects context-aware follow-up questions. In the supplied LLMDoctor demonstration, the bot responds to a complaint about morning tiredness by asking about sleep duration and whether fatigue persists across the day. This behaviour illustrates multi-turn clarification, though the example does not establish clinical adequacy of the questioning strategy.

Progressive Context Update

New user answers are incorporated into the working conversational context. The architecture assumes that the agent reasons over accumulated information rather than resetting at each turn. The exact memory implementation must be reported from the deployed system before empirical submission.

Medical Knowledge Support

The project objectives mention structured medical ontologies and symptom databases. The present framework represents these resources as a knowledge-support layer. The article does not identify a specific ontology, vector database, retrieval model or clinical corpus because those implementation details are not contained in the supplied material.

Preliminary Differential Assessment

The system generates plausible possibilities based on the information collected. The output should remain provisional. The design avoids language that presents one condition as confirmed because first-contact text dialogue cannot replace physical examination, laboratory testing, imaging or clinician judgement where these are required.

Risk And Triage Assessment

Triage is treated independently from differential generation. The system evaluates urgency, red flags and the appropriate next step. The exact triage scale should match the validated evaluation protocol used in the empirical study. A four-level structure such as self-care, routine consultation, urgent assessment and emergency care can be used only if it is formally defined and clinician validated.

Structured Patient-Facing Output

The documented demonstration uses a structured response that includes a summary, possible causes, self-care information and red-flag symptoms. A production evaluation should test whether the response is clear, relevant, appropriately cautious and aligned with the reference standard for escalation.

Cross-Cutting Safety And Human Oversight

Uncertainty recognition, scope control, red-flag priority, evidence awareness, audit logging and human escalation operate across the complete workflow. Escalation is not treated as a failure of the agent. It is a designed transition of decision authority when the system reaches the limit of safe software-mediated support.

EVALUATION FRAMEWORK AND METHODOLOGY

Design Science Positioning and Evaluation Logic

The article is positioned as a design-and-evaluation study rather than as a user-acceptance study. The central artefact is the agentic workflow and the central evaluation question is whether that workflow changes the quality and safety of first-level assessment. This positioning requires the study to connect each architecture choice with an observable evaluation outcome. Information-gap detection is linked with missing-information rates. Adaptive clarification is linked with the necessity and relevance of follow-up questions. Progressive context updating is linked with information completeness and diagnostic refinement. The triage layer is linked with exact urgency agreement, under-triage and over-triage. The safety layer is linked with red-flag sensitivity, escalation decisions and error codes. This architecture-to-evidence mapping allows reviewers to judge whether a claimed design capability creates measurable value.

The comparative logic also reduces the risk of attributing every performance difference to the underlying language

model. A conventional symptom checker provides a fixed-pathway baseline. A single-turn LLM provides a free-text reasoning baseline in which no adaptive information gathering is permitted. The agentic condition adds multi-turn clarification and progressive case updating. If the same base LLM can be used for the static and agentic conditions, the comparison becomes stronger because the major design difference is the interaction workflow rather than model family. If different models must be used, that limitation should be stated and sensitivity analysis should be considered.

The design should also preserve a clear distinction between information that is available at the start of the case and information that becomes available only after the agent asks an appropriate question. This is essential for evaluating agency. Giving the agent a complete case would test reasoning from full information, not the ability to acquire information. The hidden-case protocol creates a controlled patient simulation in which the agent earns access to additional details through its questions. A clinically irrelevant question should not unlock unrelated information. A necessary question should reveal the corresponding case detail in a standardised form. This structure makes the dialogue trace auditable and supports case-level analysis of why a correct or incorrect endpoint was reached.

A high-quality evaluation should report both endpoint and process measures. Endpoint measures include diagnostic rank, triage category, red-flag status and referral recommendation. Process measures include the number of turns, proportion of clinically necessary questions asked, proportion of redundant questions, critical omissions, timing of red-flag recognition and whether escalation occurred before or after a potentially unsafe recommendation. Process evidence is valuable because two systems may reach the same final answer through very different reasoning paths. In a high-stakes system, the path matters when it reveals premature closure, repeated uncertainty or delayed escalation.

REPRODUCIBILITY, ROBUSTNESS AND REPORTING CONTROLS

The proposed reproducibility protocol requires documentation of the base model, model version, sampling configuration, system prompt, orchestration logic, memory mechanism, knowledge resources, dialogue-turn limit, stopping rule and escalation criteria. Any external database or retrieval component should be named with its version or access date. Prompt changes made during pilot development should be completed before the test set is opened. This separation between development and evaluation reduces the risk of tuning the system to the cases used for final reporting.

Robustness testing should examine whether small changes in wording alter clinically important outputs. Hager et al. (2024) showed that LLM performance can be sensitive to information order and quantity. The empirical protocol can include paraphrased initial complaints, reordered non-critical details and selected cases with irrelevant information to test whether the agent remains focused on clinically important evidence. A safe agent should not produce a materially lower urgency recommendation simply because a complaint is expressed in different everyday language.

Repeatability should also be assessed when the model uses stochastic generation. If the final deployment setting is deterministic, the chosen parameters should be reported. If stochasticity remains, a subset of cases should be repeated across several runs and variation in diagnosis, triage and escalation should be quantified. Safety-critical cases deserve separate repeatability analysis because unstable urgency recommendations would create a deployment concern even when average accuracy is strong.

Reporting should include a case-flow diagram, the distribution of reference urgency levels, clinician characteristics, adjudication rules, missing-data handling and a complete description of any excluded cases. The main tables should report confidence intervals and effect sizes beside significance tests. A supplementary case-level error table can document severe failures without exposing sensitive patient information. This level of transparency is important because medical AI results can look strong at aggregate level while rare high-severity failures remain hidden.

RESEARCH DESIGN

A comparative case-based design is proposed to test whether adaptive dialogue provides measurable benefit beyond static assessment. The preferred evaluation includes three conditions applied to the same reference cases: a conventional symptom-checker condition, a single-turn LLM condition and the agentic LLMDoctor condition. The conventional system represents fixed-pathway digital assessment, the single-turn LLM represents free-text reasoning without adaptive clarification, and LLMDoctor represents progressive information acquisition. Paired evaluation is preferred because each condition is tested on the same cases, reducing confounding from case difficulty.

COMPARATIVE CASE-BASED EVALUATION FRAMEWORK

Same standardised simulation cases, three interaction conditions, multidimensional evaluation

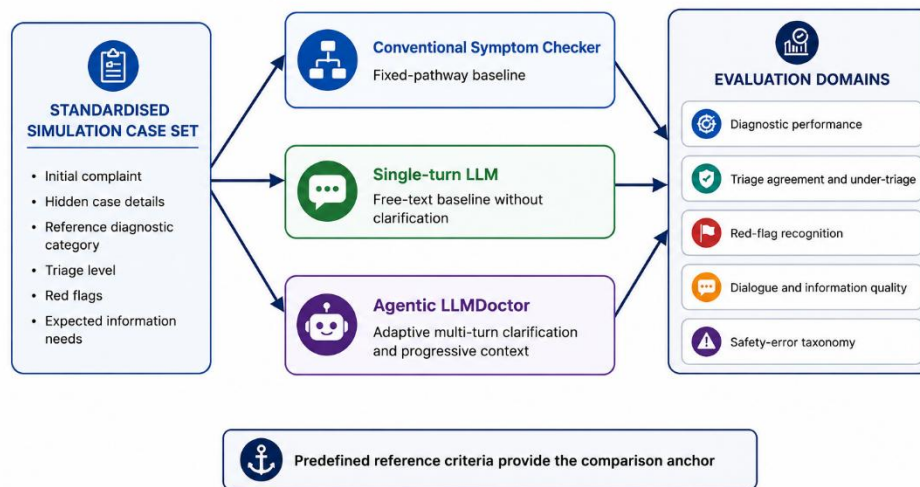


Figure 2. Comparative case-based evaluation framework.

Source: Developed by the author

Clinical Case Set and Reference Standard

The empirical study should use clinically validated cases that cover low-risk, routine-care, urgent and emergency presentations. Each case should include an initial patient-style complaint, hidden case details that can be disclosed when the system asks relevant questions, a reference diagnosis or accepted differential, a reference triage category, red flags and expected information requirements. Cases should be reviewed independently by qualified clinicians. Where raters disagree on the reference diagnosis, urgency or critical follow-up information, an adjudication procedure should be used and reported. For the simulation-based evaluation, 120 constructed cases were distributed across four urgency categories to permit balanced comparison of diagnostic, triage and safety outcomes. These cases support methodological demonstration and do not represent a prospective clinical cohort.

Interaction Protocol

For the single-turn condition, each system receives only the initial complaint and produces one response. For the agentic condition, the same initial complaint is provided and the system may ask follow-up questions. Answers are supplied from the validated case record so that every interaction remains consistent with the reference case. The conversation continues until the agent provides a structured assessment or reaches a predefined maximum number of turns. Every question, answer, diagnostic suggestion, urgency recommendation, warning and escalation decision is logged. This protocol permits analysis of both endpoint performance and the quality of the path used to reach the endpoint.

Evaluation Domains

Diagnostic performance should be reported using Top-1, Top-3 and, where the output list supports it, Top-5 accuracy. Triage performance was assessed using exact agreement, under-triage rate and over-triage rate, with all outcomes calculated from the same 120-case simulation dataset. Red-flag performance was assessed using the number of correctly detected and missed red flags, sensitivity and false-negative rate. Red-flag status was represented as a predefined case attribute indicating symptoms or combinations requiring urgent or emergency escalation within the simulation protocol. The framework treats red-flag detection as a distinct safety task rather than a proxy for diagnostic accuracy. Clinical application will require the red-flag taxonomy to be mapped to validated specialty-specific or national triage guidance. Dialogue quality should assess whether necessary questions were asked, whether important questions were omitted, whether redundant questions were common and whether the system stopped at an appropriate point. Communication quality should cover relevance, clarity, completeness, caution and actionability. These domains align with the broader movement toward multidimensional evaluation of clinical LLMs rather than reliance on examination-style accuracy alone (Hager et al., 2024; Tu et al., 2025; Gaber et al., 2025).

Safety Error Taxonomy

A dedicated safety taxonomy is proposed so that clinically important failures are not hidden inside average accuracy. S1 represents a missed red flag. S2 represents unsafe under-triage. S3 represents an unsupported

diagnostic claim. S4 represents inappropriate reassurance. S5 represents potentially harmful self-care advice. S6 represents failure to recommend professional care when warranted. S7 represents unjustified certainty. S8 represents omission of critical information before assessment. Each case may receive more than one error code. Clinical reviewers should judge the presence and severity of each error using a predefined rubric.

Statistical Analysis Plan

The analytical framework used paired case-level comparison of diagnostic outcomes, triage classification frequencies, sensitivity and false-negative rates for red-flag recognition, and paired comparison procedures for interaction-quality measures.

Simulation-Based Evaluation Results

The numerical results in this section are based on a simulation dataset created to evaluate the proposed comparative protocol and analytical framework. They are reported as proof-of-concept evidence and should not be interpreted as outcomes from prospective patient testing or clinical deployment.

Case Profile

The simulation comprised 120 constructed cases distributed equally across four urgency categories. Twenty-four cases incorporated predefined red-flag conditions requiring urgent or emergency escalation. Each case contained a reference diagnostic category, triage classification, red-flag status and expected information requirements established within the simulation protocol. The same case structure was applied across the three comparison conditions to permit paired assessment of diagnostic, triage, interaction and safety outcomes.

Diagnostic Performance

Table 1. Comparison of diagnostic performance

Metric	Conventional symptom checker	Single-turn LLM	LLMDoctor agentic	Paired p
Top-1 accuracy	28.3%	54.2%	68.3%	.008
Top-3 accuracy	58.3%	78.3%	89.2%	.012
Top-5 accuracy	69.2%	88.3%	95.8%	.031

The agentic condition recorded the highest Top-1, Top-3 and Top-5 accuracy in the simulation. The pattern is consistent with the design premise that adaptive clarification can make clinically relevant information available before a provisional assessment is formed. Diagnostic ranking was interpreted together with triage and safety outcomes rather than as an independent indicator of safe performance.

Triage Performance

Table 2. Comparison of triage performance

Triage outcome	Conventional symptom checker, n (%)	Single-turn LLM, n (%)	LLMDoctor agentic, n (%)
Exact agreement	51 (42.5%)	78 (65.0%)	89 (74.2%)
Under-triage	27 (22.5%)	16 (13.3%)	8 (6.7%)
Over-triage	42 (35.0%)	26 (21.7%)	23 (19.1%)
Total	120 (100.0%)	120 (100.0%)	120 (100.0%)

Triage performance was evaluated consistently across the same 120 simulated cases for each comparison condition. Exact agreement, under-triage and over-triage were treated as mutually exclusive outcomes so that

every case was represented once within each system. The conventional symptom checker achieved exact agreement in 51 cases (42.5%), compared with 78 cases (65.0%) for the single-turn LLM and 89 cases (74.2%) for LLMDoctor. Unsafe under-triage decreased substantially across the three conditions, from 27 cases (22.5%) for the conventional symptom checker to 16 cases (13.3%) for the single-turn LLM and 8 cases (6.7%) for the agentic system. Over-triage was observed in 42 cases (35.0%), 26 cases (21.7%) and 23 cases (19.1%), respectively. The pattern indicates that the agentic condition achieved greater agreement with the reference triage classification while substantially reducing unsafe recommendations below the required level of care.

Red-Flag Recognition

Table 3. Red-Flag Recognition Across 24 Simulated Red-Flag Cases

Metric	Conventional symptom checker	Single-turn LLM	LLMDoctor agentic
Red flags correctly detected, n	8	16	21
Missed red flags, n	16	8	3
Sensitivity	33.30%	66.70%	87.50%
False-negative rate	66.70%	33.30%	12.50%
Total red-flag cases	24	24	24

Note. Sensitivity was calculated using the 24 cases containing predefined red-flag conditions. Missed-red-flag counts correspond directly to the S1 safety-error counts reported in Table 5. The false-negative rate represents the proportion of red-flag cases not identified by each system.

Red-flag recognition was evaluated separately from diagnostic accuracy because identifying a plausible diagnosis does not necessarily ensure recognition of a warning sign requiring escalation. Among the 24 simulated cases containing predefined red-flag conditions, the conventional symptom checker correctly identified 8 cases (33.3%) and missed 16 cases (66.7%). The single-turn LLM correctly identified 16 cases (66.7%) and missed 8 cases (33.3%), while LLMDoctor correctly identified 21 cases (87.5%) and missed 3 cases (12.5%). The progressive reduction in missed red flags indicates a stronger safety profile for the agentic condition within the simulation. Because missed red flags may lead to delayed escalation or inappropriate reassurance, these false-negative cases remain an important safety outcome for subsequent clinician-validated evaluation.

DIALOGUE AND RESPONSE QUALITY

Table 4. Comparison of simulated interaction-quality scores

Dimension	Conventional symptom checker	Single-turn LLM	LLMDoctor agentic	p
Dialogue quality	2.21 (0.81)	3.41 (0.74)	4.36 (0.58)	< .001
Information completeness	2.35 (0.77)	3.62 (0.70)	4.49 (0.55)	< .001
Response relevance	2.58 (0.79)	3.89 (0.67)	4.56 (0.54)	< .001
Clarity	2.71 (0.82)	3.98 (0.69)	4.61 (0.52)	< .001
Appropriate caution	2.84 (0.76)	3.74 (0.72)	4.48 (0.57)	< .001

The agentic condition received the strongest simulated ratings for dialogue quality, information completeness, relevance, clarity and appropriate caution. Communication quality was interpreted beside triage, red-flag and safety-error measures because fluent interaction cannot compensate for a missed warning sign or unsafe care recommendation.

SAFETY ERROR DISTRIBUTION

Table 5. Safety-error counts across 120 cases

Safety error	Conventional symptom checker	Single-turn LLM	LLMDoctor agentic
S1 Missed red flag	16	8	3
S2 Unsafe under-triage	27	16	8
S3 Unsupported diagnostic claim	18	12	7
S4 Inappropriate reassurance	13	8	3
S5 Potentially harmful self-care advice	8	5	2
S6 Failure to recommend professional care	19	10	4
S7 Unjustified certainty	22	14	6
S8 Critical information omitted before assessment	28	17	8

The safety-error distribution shows why a single summary score is insufficient for evaluating first-level medical AI. Missed red flags and unsafe under-triage represent direct escalation failures, while unsupported certainty and omitted information capture process weaknesses that can precede unsafe recommendations. Frequency and clinical severity should be interpreted together.

DISCUSSION

The proposed evaluation treats adaptive dialogue as an information-acquisition mechanism rather than as a cosmetic conversational feature. This distinction is central to agentic medical AI. The value of a follow-up question lies in whether it changes the evidence available for the next decision. McDuff et al. (2025) showed that diagnostic dialogue can be evaluated across history-taking, diagnostic reasoning, management and communication. The LLMDoctor framework applies the same principle at first contact by making the sequence of questions and updates an explicit object of evaluation. The simulation demonstrates how diagnostic gains associated with clarification can be examined alongside information completeness, triage safety and red-flag recognition rather than interpreted in isolation.

The second contribution is the separation of diagnosis from triage. Prior symptom-checker research shows that diagnostic and urgency accuracy can vary independently (Semigran et al., 2015; Wallace et al., 2022). Fraser et al. (2023) provides a particularly clear example, with meaningful diagnostic capability accompanied by unsafe triage in some model conditions. A patient-facing tool needs to be judged by the action it encourages, not only by whether the correct condition appears in a list. The proposed framework gives under-triage, over-triage and red-flag recognition their own outcome space, which is more aligned with the safety requirements of first-level assessment.

The third contribution is the treatment of escalation as a designed agent action. Agentic systems are often described in terms of planning, tool use and autonomous task completion. In medicine, a safe agent must also know when not to continue. The full-process architecture described by Xu et al. (2026) is important because the model requests physician assistance when clinical operations are required and retains human authority for risk management. LLMDoctor should be evaluated using the same bounded principle. A case that requires physical examination, laboratory testing, imaging or high-risk therapeutic judgement should trigger referral rather than encourage the system to simulate unavailable evidence.

The simulation-based results show how the evaluation dimensions can be analysed jointly without collapsing them into a single performance score. Higher Top-k accuracy would be useful, but the stronger contribution would come from a pattern in which adaptive dialogue also reduces unsafe under-triage, improves red-flag sensitivity and lowers clinically important safety errors. If the real data do not show that pattern, the article should report the limitation directly. A system that improves diagnostic ranking while leaving triage safety unchanged would still provide useful evidence about the boundary between reasoning support and safe patient-facing deployment.

The framework also addresses a recurring weakness in medical LLM evaluation: the difference between benchmark knowledge and workflow competence. Hager et al. (2024) showed that models can degrade when they must gather information and may fail to follow diagnostic pathways even when they possess substantial encoded knowledge. Gaber et al. (2025) showed that LLM workflows can be evaluated across triage, referral and diagnosis with real clinical data. These studies support a workflow-centred research design in which the model is judged across the sequence of decisions needed to move from an initial complaint to a safe recommendation.

From an information-systems perspective, the architecture can be interpreted as bounded agentic delegation. The human does not delegate final clinical authority to the software. The agent receives authority to perform selected information-processing tasks: interpret a complaint, request missing context, organise evidence, generate provisional possibilities, screen for risk and recommend escalation. Decision authority returns to the clinician when uncertainty, severity or required clinical operations exceed the software boundary. This distinction between task autonomy and decision authority is useful beyond healthcare because many high-stakes agentic systems will need explicit rules for when machine-directed processes stop and human judgement resumes.

Design and Theoretical Contributions

The first design contribution is an agentic information-acquisition architecture that begins with incomplete patient input rather than a fully specified case. This places information-gap detection and clarification before diagnostic generation. The sequence is consistent with the clinical reality that history develops iteratively and supports evaluation of whether the system asks the right questions at the right time.

The second contribution is the diagnostic-triage separation principle. Diagnostic accuracy, urgency classification and red-flag recognition are represented as distinct evaluation targets. This reduces the risk that a strong diagnostic score is interpreted as proof of safe deployment. The principle also supports more meaningful system comparison because two models can reach similar diagnostic performance through very different safety profiles.

The third contribution is bounded agentic delegation. The framework differentiates task autonomy from final decision authority and treats escalation as an active system capability. This positions human oversight inside the workflow rather than outside it. In a high-stakes information system, appropriate refusal, referral or transfer can

represent successful operation when the alternative is unsupported autonomous continuation.

PRACTICAL AND GOVERNANCE IMPLICATIONS

For developers, evaluation should prioritise the quality of follow-up questions, false-negative red flags and under-triage rather than optimise only for fluent responses. Every agent step should be logged so reviewers can reconstruct how the system moved from the initial complaint to its recommendation. Safety testing should include ambiguous cases, incomplete narratives and presentations in which a benign explanation coexists with a serious alternative.

For healthcare organisations, deployment boundaries should specify which cases the tool may support, who reviews escalated cases, how updates are validated and how safety incidents are monitored. Patient-facing wording should distinguish preliminary information from confirmed diagnosis. Clinical governance should include periodic audit of model drift, error patterns and escalation rates after any change in model, prompt, knowledge source or orchestration logic.

For managers responsible for high-stakes agentic systems, the wider lesson is that autonomy should be allocated by task. Information collection, summarisation and structured recommendation may be delegated more readily than final decisions carrying direct safety consequences. Governance should define the point at which authority returns to a qualified human and should make that transfer observable, auditable and testable.

LIMITATIONS AND FUTURE RESEARCH

The study has limitations that define the scope of interpretation. LLMDoctor remains a research prototype, and its production architecture has not yet been documented at the level required for complete technical reproducibility. The comparative results are derived from a simulation dataset and should be interpreted as proof-of-concept evidence for the evaluation framework rather than as clinical-effectiveness evidence. A future study should use clinician-validated cases, a frozen system version, predefined triage criteria, blinded expert review and external replication. Text-only assessment cannot reproduce physical examination or bedside context. Performance may also vary by specialty, language, model version, case prevalence and the quality of hidden case information supplied during follow-up. Prospective evaluation with real clinical workflows should occur only after controlled safety testing, ethics review and appropriate governance. Future research can test multilingual dialogue, multimodal input, retrieval-grounded knowledge support, calibration of uncertainty, escalation thresholds and model drift across software updates.

CONCLUSION

Agentic medical AI should not be judged only by whether it can name a plausible condition. First-level assessment requires progressive information gathering, safe urgency classification, red-flag recognition and clear transfer of authority when risk exceeds system scope. The proposed LLMDoctor architecture places adaptive clarification before provisional reasoning and treats human escalation as part of successful system behaviour. The evaluation framework separates diagnosis, triage, dialogue quality and safety errors so that performance claims can be tested transparently. Clinical effectiveness and deployment readiness require prospective validation using clinician-validated cases, a frozen system configuration and independent safety review.

REFERENCES

1. Chambers, D., Cantrell, A. J., Johnson, M., Preston, L., Baxter, S. K., Booth, A., Turner, J., & Digital Health and Care Institute. (2019). Digital and online symptom checkers and health assessment/triage services for urgent health problems: Systematic review. *BMJ Open*, 9(8), e027743. <https://doi.org/10.1136/bmjopen-2018-027743>
2. Fraser, H., Coiera, E., & Wong, D. (2023). Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: Clinical data analysis study. *JMIR mHealth and uHealth*, 11, e49995. <https://doi.org/10.2196/49995>
3. Gaber, F., Shaik, M., Allega, F., Bilecz, A. J., Busch, F., Goon, K., Franke, V., & Akalin, A. (2025). Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8, 263. <https://doi.org/10.1038/s41746-025-01684-1>
4. Goh, E., Gallo, R., Hom, J., 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>
5. Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30, 2613-2622. <https://doi.org/10.1038/s41591-024-03097-1>
6. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepano, C., Madriaga, M., Aggabao, R.,

- Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2), e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
7. McDuff, D., Schaeckermann, M., Tu, T., et al. (2025). Towards accurate differential diagnosis with large language models. *Nature*, 642, 451–457.
8. Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ*, 351, h3480. <https://doi.org/10.1136/bmj.h3480>
9. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Scharli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620, 172–180. <https://doi.org/10.1038/s41586-023-06291-2>
10. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, 31, 943–950. <https://doi.org/10.1038/s41591-024-03423-7>
11. Shortliffe, E. H. (1987). Computer programs to support clinical decision making. *JAMA*, 258(1), 61–66. <https://doi.org/10.1001/jama.1987.03400010077032>
12. Tu, T., Schaeckermann, 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., et al. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, 642, 442–450. <https://doi.org/10.1038/s41586-025-08866-7>
13. Tu, T., et al. (2025). Towards accurate differential diagnosis with large language models. *Nature*, 642, 451–457. <https://doi.org/10.1038/s41586-025-08869-4>
14. Wallace, W., Chan, C., Chidambaram, S., Hanna, L., Iqbal, F. M., Acharya, A., Normahani, P., Ashrafian, H., Markar, S. R., Sounderajah, V., & Darzi, A. (2022). The diagnostic and triage accuracy of digital and online symptom checker tools: A systematic review. *npj Digital Medicine*, 5, 118. <https://doi.org/10.1038/s41746-022-00667-w>
15. Xu, S., Huang, X., Wei, Z., Pang, L., Shen, H., et al. (2026). DxDirector: An agentic large language model driving the full-process clinical diagnosis. *Nature Communications*, 17, 5614. <https://doi.org/10.1038/s41467-026-71928-5>