CLARITY: Clinical Assistant for Routing, Inference, and Triage

Vladimir Shaposhnikov[*,1,2] **Aleksandr Nesterov**[*,1] **Ilia Kopanichuk**[*,1,3]
Ivan Bakulin[1,3] **Egor Zhelvakov**[1] **Ruslan Abramov**[4]
Ekaterina Tsapieva[5] **Iaroslav Bespalov**[1] **Dmitry V. Dylov**[1] **Ivan Oseledets**[1]
[1]AIRI [2]Skoltech [3]MIPT [4]SberMedAI [5]Sber
[*] Equal contribution
Russia
{shaposhnikov, nesterov, kopanichuk}@airi.net

Abstract

We present CLARITY (Clinical Assistant for Routing, Inference and Triage), an AI-driven platform designed to facilitate patient-to-specialist routing, clinical consultations, and severity assessment of patient conditions. Its hybrid architecture combines a Finite State Machine (FSM) for structured dialogue flows with collaborative agents that employ Large Language Model (LLM) to analyze symptoms and prioritize referrals to appropriate specialists. Built on a modular microservices framework, CLARITY ensures safe, efficient, and robust performance, flexible and readily scalable to meet the demands of existing workflows and IT solutions in healthcare.

We report integration of our clinical assistant into a large-scale national interhospital platform, with more than 55,000 content-rich user dialogues completed within the two months of deployment, 2,500 of which were expert-annotated for subsequent validation. The validation results show that CLARITY surpasses human-level performance in terms of the first-attempt routing precision, naturally requiring up to 3 times shorter duration of the consultation than with a human.

1 Introduction

The integration of LLMs into healthcare could truly transform the industry. These models demonstrate potential to improve diagnostic accuracy, optimize clinical workflows, and enhance the overall experience for patients and physicians. Dialog systems powered by LLMs are particularly promising, as they are capable of automating routine tasks, generating initial diagnostic hypotheses to optimize access to care and reduce waiting times (Tu et al., 2024; Bao et al., 2023; Chen et al., 2024; Tripathi et al., 2024; Meng et al., 2024).

Medical dialogue systems are designed to facilitate clinical consultations, assist healthcare providers, and connect patients with proper expertise. Unlike general-purpose conversational AI, these systems must be able to handle complex diagnostic hypotheses reasoning, extended back-and-forth interactions, and domain-specific medical knowledge. Although recent advances in LLMs have significantly improved their conversational abilities, they are still fundamentally flawed for real-world adoption in healthcare.

Despite all the efforts, modern patient routing approaches often involve unnecessary delays. Patients typically consult general practitioners (GPs) as an intermediary step before being referred to specialized experts, which increases wait times and complicates timely access to care. Automated assessment of patient conditions with a generated diagnostic hypotheses could significantly streamline this process, benefiting both patients and healthcare providers (Umerenkov et al., 2025; Nazi and Peng, 2024). This optimization also promises substantial economic advantages, potentially reducing healthcare costs and improving the allocation of resources.

The potential economic impact extends to the rapidly growing telemedicine market. While telemedicine offers expanded access to care, high consultation costs remain a significant barrier. Automating certain GP functions through LLM-powered systems offers a pathway to more affordable and accessible telemedicine services (Downes et al., 2015; Waller and Stotler, 2018). Similarly, improvements in online appointment scheduling systems, often hindered by user interface challenges and integration issues, could further enhance patient experience and reduce administrative burdens (Ala et al., 2023; Gupta and Denton, 2008).

Several specialized LLMs, including AMIE, DISC-MedLLM, CDD, and CoD (Tu et al., 2024; Bao et al., 2023; Chen et al., 2024), demonstrate the potential of this technology. While systems like AMIE show high diagnostic hypotheses accuracy, challenges remain regarding scalability and

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1805–1821
November 4-9, 2025 ©2025 Association for Computational Linguistics

reliance on specialized datasets. Others, such as DISC-MedLLM, prioritize empathetic patient interaction, while CDD focuses on structured symptom questioning. However, the field lacks unified standards for training, evaluation, and deployment, hindering model comparison and regulatory approval (Bedi et al., 2024; Singhal et al., 2023).

Despite these advancements, critical challenges persist. LLMs may generate inaccurate or fabricated information ("hallucinations")(Huang et al., 2024; Tonmoy et al., 2024; Pal et al., 2023), a particularly dangerous risk in medicine. Current models also struggle to reliably identify life-threatening conditions and may exhibit inconsistencies in patient interaction, potentially affecting trust. These issues necessitate robust solutions to ensure safe and effective integration of LLMs into clinical practice (Zhou et al., 2024a; Yang et al., 2024; Shi et al., 2024).

To address these challenges, this article introduces CLARITY, a hybrid system integrating the strengths of rule-based and LLM approaches. CLARITY is designed to provide a flexible, controllable, and accurate medical dialog system suitable for real-world applications. By leveraging FSM, a microservices architecture, and specialized datasets, CLARITY facilitates structured, context-aware patient interactions, enhancing diagnostic hypotheses reliability, enabling real-time critical condition recognition, and providing personalized patient guidance.

2 System Architecture

The CLARITY system is designed as a hybrid architecture, integrating an FSM for the management of dialogue with a microservices architecture for the processing of requests.

The architecture consists of the following core components:

- **FSM** governs the dialogue flow by managing states and transitions based on user input and context. **Text generation services** are responsible for a natural language generation and complex query processing. In our deployment, the LLM used for generation and analysis is GigaChat[1].
- **Decision making services** are responsible for input text classification and decision-making.
- **Microservice architecture** ensures modularity and scalability by isolating individual components; real-time performance further depends on minimizing inter-service latency and optimizing orchestration.

[1] https://giga.chat/

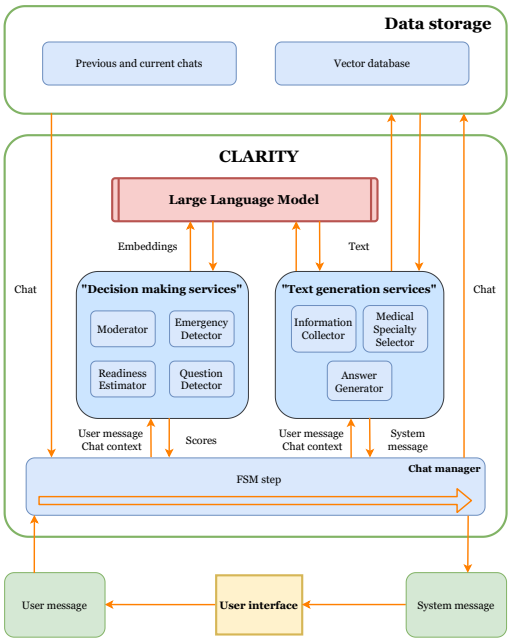


Figure 1: The architecture of the CLARITY system.

Figure 1 illustrates the architecture of the CLARITY system, showing how its core components interact. The FSM serves as the backbone of dialogue management, ensuring structured and predictable interactions, while text generation and decision making services provide classification and response generation capabilities, respectively.

2.1 Chat Manager

The Chat Manager in the CLARITY system is implemented as a FSM, tailored to specific requirements of the medical domain. Its primary goals are to ensure structured dialogue flow, maintain patient safety, and handle sequential information gathering.

States. The FSM has six dialogue contexts:

- *Initialization* - greets the user and begins collecting initial complaints;
- *Information Collection* - coordinates the information - gathering module for symptoms and medical history;
- *Diagnostic Hypothesis & Routing* - processes collected data to generate preliminary diagnostic hypotheses, select medical specialty, and formulate recommendations;
- *Moderation* - prevents unsafe dialogue scenarios;
- *Emergency* - executes predefined protocols for urgent cases;
- *Free Dialogue* - handles non-standard requests via the answer-generation module.

Transitions. Movement between states depends

on user input, dialogue history, and contextual signals (see Appx. E.3); decisions are driven by outputs from dedicated decision-making services and the FSM's internal logic.

Output Functions. Each state defines its own system responses, encompassing natural-language generation by LLMs and calls to internal microservices (see Appx. E.4).

For a more detailed formal description of the FSM and its components, please refer to Appendix E.

By combining the structured control of the FSM with the adaptability of LLMs, the Chat Manager achieves a balance between *reliability* and *flexibility*. First, **predictability and safety** are ensured by the FSM's ability to control the dialogue flow and enforce strict transitions between states. Second, the FSM's modular design ensures **scalability**, making it easy to integrate new states, transitions, and services as needed. Finally, the FSM is specifically tailored to meet **medical domain requirements**. Moderation state, critical situation handling, and strict control mechanisms ensure the system adheres to the high safety and reliability standards demanded in this field.

2.2 FSM Scalability

CLARITY's FSM addresses scalability concerns through hierarchical decomposition designed to prevent architectural brittleness as clinical scope expands:

Multi-level FSM Design. The current implementation employs logical decomposition where a global FSM graph describes overall system behavior and manages high-level transitions between initialization, information collection, diagnostic hypotheses, moderation, and emergency handling states. Simultaneously, local FSM graphs govern the logic within individual medical scenarios and specialty-specific workflows, creating a hierarchical structure that naturally distributes complexity across multiple manageable components.

Modular Expansion Strategy. New medical domains can be integrated by:

- Defining domain-specific local FSM graphs that handle specialty workflows.
- Adding appropriate transition points in the global FSM to route users to new domains.
- Implementing corresponding microservices that provide domain-specific processing capabilities.

This approach ensures system complexity grows linearly rather than exponentially with clinical area additions.

Graph-based Storage and Analysis. The FSM is stored as a graph structure in memory, enabling development of semi-automated analysis tools for: redundant cycle detection, unreachable state identification, conflicting transition condition analysis, and other architectural consistency checks that support long-term maintainability.

2.3 Services

2.3.1 Moderator

The moderator checks whether a given request or model response contains a prohibited topic. A separate moderation skill is needed for two reasons. First, it is essential to limit sensitive topics, aggression, and hatred in conversations. Second, default moderation systems often cannot handle medical data due to compliance requirements. Moderation is formulated as a binary classification problem: 0 (no prohibited topics) or 1 (prohibited topics detected). The proposed algorithm uses a nested blacklist of prohibited words specific to each topic. Performance is evaluated using the F_1-score.

2.3.2 Emergency Detector

The emergency detector checks whether a user is in need for emergency medical care. It is a binary classification problem: 0 (no emergency) or 1 (emergency). The algorithm processes input chat W into a text string of user and model messages via concatenation. The function σ is computed as:

$$\sigma_{tr}(W) = \text{HGB}\Big(\text{PCA}\big(\text{concat}(\text{tfidf}(W), \text{OHE}(W), \text{LLM}(W)), n_c\big)\Big) > t \quad (1)$$

where t is the threshold value, HGB is a hist-based gradient boosting, and $n_c = 423$ is the number of components for PCA. We enhance the concatenation function by including a one-hot-encoder (OHE) for critical words and a binary value indicating whether the LLM model considers the condition critical. Performance is evaluated using F_1-score and false positive rate (FPR).

2.3.3 Readiness Estimator

The readiness estimator determines the appropriate time to transition from the information collection phase to the referral stage. It mitigates the risk of premature or delayed transitions, improving user experience and system accuracy. The module is trained on a corpus of 2,500 medical dialogues,

using TF-IDF and contextual embeddings for feature extraction. Performance is evaluated using the mean absolute percentage error (MAPE) and the F_1-score.

2.3.4 Question Detector

The question detector is responsible for the binary classification of user messages with the objective of determining whether a message constitutes a clarifying question. This module ensures flexibility in the system's workflow for collecting medical history and complaints. Performance is evaluated using F_1-score.

2.3.5 Information Collector

The information collector module gathers comprehensive information on patient complaints and medical history. It includes a question generator based on a language model, a dialogue search system, and a question relevance classifier (see Appx. F).

1. The module begins by receiving the user's message and dialogue history.
2. It then searches the historical dialogue vector database; if a similar case is found (cosine similarity > 0.965), previously generated questions are reused.
3. If no match is identified, the system uses a domain-specific prompt to generate five new candidate questions with the LLM.
4. To avoid redundancy, each candidate is compared to prior dialogue questions, and those with cosine similarity above 0.86 are discarded.
5. Remaining candidates are passed through a relevance classifier trained on 2,500 labeled examples; the classifier achieves 0.84 precision on the "relevant" class.
6. The most relevant question is selected and presented to the user. This architecture ensures both contextual accuracy and low response latency.

2.3.6 Medical Specialty Selector

The medical specialty selector is a core module that generates candidate diagnoses, identifies appropriate medical specialists, and provides supporting explanations. The process includes three sequential stages (Appx. G):

1. **Diagnostic Hypotheses Generation:** Using the patient's complaints and medical history, the system queries the LLM to generate N possible diagnostic hypotheses. If fewer than N are returned, the request is repeated to ensure sufficient coverage.

Doc: What's bothering you?
Pt: I have a headache.
Doc: Where exactly is the pain located?
Pt: The back of my head.
Doc: Are you experiencing any other symptoms, such as nausea or vomiting?
Pt: No.
Doc: How intense is the pain?
Pt: 5 out of 10.
Doc:
Cervicogenic headache – *General practitioner*. Pain in the back of the head may be related to cervical-spine issues.
Cervical osteochondrosis – *Neurologist*. Neck pathology may provoke occipital pain.
Doc: Is everything clear? Feel free to ask questions!

Figure 2: Dialogue example illustrating the *Transparency* scenario. Full dialogue examples for *Critical*, *Safety*, and *Adaptability* are provided in Appendix C, Table 4.

2. **Specialist Selection:** For each diagnostic hypotheses, a parallel LLM call determines the most relevant medical specialist (e.g., cardiologist for cardiovascular symptoms, gastroenterologist for digestive issues), ensuring that each condition is matched with an appropriate expert.

3. **Explanation Generation:** The system produces a short explanation for each diagnosis-specialist pair, including a description of the condition, reasoning behind the diagnostic hypotheses, and rationale for the referral. These explanations aim to improve patient understanding and transparency.

The module is optimized via parallel processing and selective regeneration of incomplete outputs, which significantly reduces latency. Its performance is evaluated using pairwise precision and recall, comparing system recommendations to expert annotations.

2.3.7 Answer Generator

This module allows for the management of arbitrary user queries and open dialogues within the context of medical consultations. The module is activated in two scenarios:

1. A deviation from the dialogue script may occur when the user poses a question that falls outside the scope of the current predefined dialogue flow.

2. Once the standard dialogue script has been completed and all necessary information has been collected, as well as a preliminary diagnostic hypotheses provided, the dialogue can be considered complete. In this instance, the open-dialogue module permits the user to pose supplementary queries, seek elucidation, or request general health-related

Table 1: Services metrics

	Value
Transparency	Information collector (P = 84 %, $f_{\text{rep}} = 0$); readiness estimator (MAPE = 22 %); medical-specialty selector (R@3 = 96 %)
Critical	Emergency detector (Precision = 72 %, Recall = 49 %, F_1 = 59 %, FPR = 2 %)
Safety	Moderator (F_1 = 95 %)
Adaptability	Question detector (F_1 = 94 %)

Table 2: Pilot statistics

	Value
Transparency	55 k+ dialogues; 3× faster than human; 80 % "friendly"; 32.4 % conversion
Critical	7.4 % dialogues flagged critical
Safety	3.4 % non-target dialogues
Adaptability	6.5 % non-standard dialogues

information.

The functionality of this module is based on the capabilities of LLMs, which are able to generate coherent and contextually relevant responses within a dialogue. In order to adapt the LLMs to the medical domain, a specialised system prompt is employed. This sets the context of the interaction, constrains the knowledge domain to medical topics and guides the model towards generating responses that are ethically sound and informative. This module facilitates smooth transitions between predefined dialogue scripts and open-ended discussions, ensuring the system's flexibility and its capacity to respond effectively to a diverse range of user queries.

3 Experiments And Results

In this section, we present a comprehensive evaluation of CLARITY's performance across its core components. The experiments were conducted on unique datasets, annotated by licensed doctors and medical experts to ensure high-quality ground truth labels. We evaluate the system using a combination of standard metrics (e.g., precision, recall, F_1-score) and custom metrics tailored to the medical domain (e.g., pairwise precision and recall for specialist selection). Additionally, we discuss the results of pilot studies, which demonstrate the system's practical applicability in real-world scenarios.

Across the safety-critical modules, the *Moderator* achieved an F_1-score of 0.95 with a false-positive rate below 1.5 %, indicating that the system rarely blocks benign content while effectively filtering prohibited topics. The *Emergency Detector* demonstrated a precision of 0.72, recall of 0.49, and an F_1-score of 0.59 at an operating point that keeps the false-positive rate below 0.02. This threshold balances the risk of missed emergencies against the clinical burden of false alarms. Together, these figures confirm the reliability of the safety layer that gates downstream actions.

The remaining components likewise show strong offline performance. The *Readiness Estimator* attains an F_1 of 0.78 and a mean absolute percentage error of 22 % when forecasting the anamnesis phase, allowing the dialogue manager to time interventions accurately. The *Question Detector* reaches macro-$F_1 = 0.94$ ($\text{recall}_{\text{pos}} = 0.87$, $\text{precision}_{\text{neg}} = 0.99$), ensuring that clarification prompts are issued only when needed. The *Information Collector* maintains 0.84 precision for question relevance while responding in 5 s on average, with 20 % of prompts served from cache. Finally, the *Medical-Specialty Selector* ranks its recommendations effectively: its top suggestion is correct in 80 % of cases, and the top two cover 95 %, cor-responding to Precision@1 = 77 % and Recall@3 = 96 %. These results collectively demonstrate that each subsystem meets the accuracy and latency requirements necessary for safe real-world deployment.

A comprehensive breakdown of datasets, model architectures and training protocols is available in Appendix D, while Tables 1 and 2 collate the key service-level metrics and pilot statistics across all four scenarios. Figure 2 reproduces a complete *Transparency* consultation, full transcripts for the remaining cases - *Critical*, *Safety*, and *Adaptability* - are available in Appendix C (Table 4).

3.1 First Pilot Study: Understanding User Behavior And Preferences

A first pilot study was conducted at the end of the fourth quarter of 2024 on an active dialogue platform with the aim of evaluating the effectiveness of CLARITY in real-world conditions, as well as identifying user preferences and potential avenues for improvement. The study encompassed a comprehensive assessment of socio-demographic indicators and a detailed analysis of user preferences. Moreover, the pilot incorporated an in-depth examination of user feedback on the current functionality of the system, offering invaluable insights into the usability and user experience.

The pilot studies were conducted on a nationwide inter-hospital telehealth platform where CLARITY was fully integrated into the standard user onboarding and consultation workflow; users

interacted with the system in their usual telehealth environment (both new registrants and returning users). Participants received clear onboarding notifications about the AI assistant and could request a hand-off to a human specialist at any time.

A total of 64% of users who completed the onboarding process for the services in question proceeded to initiate a dialogue. Of the 1,534 initiated consultations, 62.13% resulted in recommendations from a specialist and a diagnostic hypothesis. The mean time taken for an appointment to conclude, including the provision of recommendations, was 2 minutes 13 seconds. This is considerably less than the time typically allocated for an in-person consultation with a general medical practitioner.

The pilot study resulted in 785 distinct preliminary diagnostic hypotheses, with referrals made to 70 physicians across various specialties. The most frequently recommended specialists were neurologists, otolaryngologists, and gastroenterologists, which is indicative of the relevance of these specialties within the scope of the pilot study (Appendix K). Neurological, renal, and gastrointestinal disorders were the most commonly diagnosed conditions. A comparison of the socio-demographic characteristics of the participants and those of the overall user base of the platform indicated a correspondence between the two groups in terms of both age and gender (Appendix H and Appendix I).

A total of 18.2% of users indicated that they found the consultation to be fully useful, while 54.5% of respondents rated it as partially useful. However, 31% of users reported that the number of questions asked was insufficient, potentially pointing to limitations in audience targeting or the depth of the dialogue process. On a positive note, 80% of users rated the interaction as friendly, which is a key factor in maintaining user engagement and satisfaction (Appendix J).

3.2 Second Pilot Study: Focus On Real-World Application

A second pilot study was conducted to evaluate the real-world application of CLARITY in a larger-scale setting. This study focused less on audience behavior and more on the practical implementation and outcomes of the system. During the pilot period, a total of 55,856 dialogues were conducted. The dialogue funnel analysis revealed a 92.3% conversion rate to the next step after dialogue initiation, with 54.7% of dialogues resulting in diagnostic hypotheses and a referral to a physician. In all cases, CLARITY performed the initial triage and routed users toward the appropriate channel. For appointment conversions, "online" consultations (telemedicine appointments booked and conducted via the platform) converted at 26.8%, whereas "offline" consultations (in-person visits at clinics or hospitals) converted at 5.6%.

The second pilot study demonstrated the scalability and practical utility of CLARITY in real-world conditions, with a high rate of dialogue progression and a significant proportion of cases leading to actionable medical outcomes. The conversion rates to both online and offline consultations further underscore the system's effectiveness in facilitating user engagement and follow-through.

4 Discussion

The CLARITY system demonstrates the effectiveness of a hybrid approach to medical dialogue systems, combining structured FSM-based dialogue management with the flexibility of large language models. Our experimental results validate this architecture's advantages while revealing important insights about AI deployment in healthcare settings.

Clinical Accuracy and Routing Efficiency. The system achieved remarkable performance in specialist routing, with 77% precision for first recommendations and 96% recall for top-3 recommendations. This surpasses typical human-level performance in initial routing while requiring only one-third of the usual consultation time (mean 2m 13s vs $\approx$ 6-7 minutes for GP practice). These results are particularly significant given current healthcare bottlenecks. The traditional model, where patients must first consult general practitioners for specialist referrals, creates substantial delays in accessing specialized care. CLARITY demonstrates that AI systems can effectively assume this initial triage function, potentially accelerating patient access to appropriate specialists while reducing the burden on primary care providers.

Critical Condition Recognition The emergency detection figures confirm an **operationally acceptable performance** under the current stakeholder-defined false-alarm budget, while highlighting the need for future recall improvements. In real-world deployment, the system identified 7.4% of cases as requiring urgent intervention, aligning with emergency department triage statistics. This achieve-

ment is particularly noteworthy as existing medical AI systems often struggle with reliable critical condition identification. CLARITY's hybrid architecture, combining FSM rules with LLM flexibility, provides a more dependable approach to detecting potentially dangerous situations.

Engagement Metrics. The 32.4% conversion rate to actual appointments indicates strong user trust and clinical relevance, substantially higher than typical digital platform metrics.

User Experience. The 80% positive interaction ratings demonstrate the system's ability to maintain empathetic and professional communication while adhering to clinical protocols. Operational Efficiency: The mean consultation time of 2 minutes 13 seconds represents a significant improvement over traditional consultations, without compromising diagnostic hypotheses accuracy.

Dialogue Control. The readiness estimator's F1-score of 78% and the question detector's accuracy of 94% demonstrate CLARITY's sophisticated dialogue management capabilities. The near-zero repetition rate (frep = 0%) in information collection confirms efficient information gathering without redundancy. Only 3.4% non-target dialogues indicates exceptional conversation management, addressing a common limitation of pure LLM-based systems.

These results suggest that successful medical AI systems require more than just powerful language models – they call for carefully designed hybrid pipelines with the strengths of different approaches of strict safety and reliability standards. CLARITY's performance metrics demonstrate that such hybrid systems can achieve both the high accuracy and the appreciation of users, creating a viable path for AI integration in a clinical setting.

5 Conclusion

In conclusion, CLARITY advances the field of medical dialogue systems by successfully addressing the major challenges that have hindered the widespread adoption of AI-powered conversational agents in healthcare. By seamlessly integrating structured reasoning, multi-agent collaboration, and robust safety measures, CLARITY sets a new standard for reliable, efficient, and user-centric medical assistance.

The system's strong performance in real-world settings, coupled with the positive user feedback, showcases its potential to facilitate the delivery of healthcare services. The insights gained from the development and evaluation of CLARITY will undoubtedly inspire and guide future efforts in this field, ultimately leading to the development of more accessible, efficient, and personalized healthcare assistants.

6 Contribution Statement

Vladimir Shaposhnikov, **Aleksandr Nesterov**, and **Ilia Kopanichuk** contributed equally: they formulated the research, designed and ran the experiments, performed the analysis, and drafted the core manuscript text. **Ekaterina Tsapieva** and **Ruslan Abramov** led the productization effort and guided the introduction of the solution into a real-world workflow. **Ivan Bakulin** and **Egor Zhelvakov** were responsible for technical deployment and production integration. **Iaroslav Bespalov**, **Dmitry V. Dylov** and **Ivan Oseledets** are co-senior authors who supervised the project and analysed the results. All authors took part in preparing the final manuscript.

7 Limitations

Despite the strengths of the CLARITY system, several limitations should be acknowledged:

- **Limited interpretability of LLM outputs.** While the system provides structured responses and intermediate explanations, it does not offer full introspection into model reasoning. Improving interpretability is an ongoing research challenge in medical AI. In future iterations, we plan to incorporate chain-of-thought tracing and example-based rationales to make the decision process more transparent for clinical users.
- **Model identity disclosure.** The system uses GigaChat as the foundation model for generation and analysis. While this work is designed to be LLM-agnostic, all reported results were obtained with GigaChat under the production configuration used in deployment.
- **Generalization to rare or atypical conditions.** The system has primarily been validated on common and well-represented clinical cases, as these reflect the majority of real-world usage scenarios where CLARITY is currently deployed. Evaluation on rare diseases, complex multi-morbidity cases, and ambiguous complaints remains limited. We plan to address this in future development by expanding the training dataset and annotation pipeline to include synthetic and curated edge cases, thereby supporting broader and more robust generalization.

- **Dialogue depth and coverage.** In pilot studies, 31% of users reported that the consultation asked "too few" questions. A formal root-cause analysis has not yet been performed; it is scheduled for Q4 2025 and will be led by an interdisciplinary team of clinicians and product specialists. The study will review a stratified sample of under-questioned dialogs to identify failure modes (e.g. premature readiness cut-off or vague initial complaints) and will feed its findings into the next iteration of the Readiness Estimator and Information Collector. Planned mitigations include (i) adding dialogue-length and answer-entropy features to the estimator and (ii) augmenting the training cache with synthetic edge-case dialogs. Our target is to lift the "question sufficiency" satisfaction rate above 85% while keeping the average consultation time unchanged.
- **Time–Satisfaction Mismatch and Explanation Factuality Limit** We acknowledge that shorter consultations do not necessarily translate into higher user satisfaction or trust: although median time decreased, many users still rated interactions as only partially useful, and we did not establish a causal link between duration, perceived usefulness, and trust. In addition, while we evaluated routing accuracy, we did not conduct a dedicated factuality (hallucination) audit of generated explanations. Future iterations will include targeted user-trust studies and clinician-in-the-loop audits of explanation correctness.
- **Lack of clinical outcome validation.** Although diagnostic hypotheses and referral suggestions were quantitatively evaluated, long-term patient outcomes were not tracked. This is due to regulatory and data privacy limitations in pilot deployments. As part of ongoing collaborations with clinical partners, we plan to conduct retrospective chart reviews and, where possible, prospective outcome-linked evaluations in future studies.
- **Absence of active learning pipeline.** Currently, CLARITY does not implement real-time self-updating mechanism or active learning. This design choice reflects critical regulatory and ethical constraints guiding the deployment of medical AI. Medical AI systems must pass stringent validation before algorithmic changes can be implemented, making real-time updates impossible from a regulatory standpoint. Furthermore, self-updating models based on patient interactions raise privacy concerns and may lead to unpredictable behavioural changes that undermine confidence in clinical decision-making. As future work, we are developing a controlled active learning pipeline that will operate under strict supervision and include offline batch learning cycles with mandatory clinical validation, explicit patient consent for the use of anonymised data, and differential privacy techniques.
- **Triage performance metrics.** The emergency-detection module is currently operated at a threshold chosen by clinical stakeholders to keep the false-positive rate below 0.02, because every alert triggers an on-call physician review and may result in ambulance dispatch. This conservative setting yields precision 0.72, recall 0.49 and F_1 = 0.59 on the expert-annotated test set. *Operational rationale:* in our deployment scenario($\approx$ 7% of dialogues genuinely critical) a higher recall would double the number of alerts and overwhelm human triage resources, whereas missed cases remain subject to the platform's standard "red-flag" questionnaire shown to all users.
  Planned mitigation. A two-stage cascade is under development: (i) a soft threshold t=0.05 to maximise sensitivity, followed by (ii) an LLM self-consistency vote (three parallel prompts, majority $\geq 2/3$) that filters obvious false alarms. The final choice of operating point will depend on a governance-board review of workload implications once additional deployment data are available. All prospective changes will be audited against the same expert-annotated test set to ensure that patient safety is never traded for convenience.
- **Partial reproducibility.** We are committed to open-sourcing all available components of our system. The code will be publicly released at: https://github.com/AnonymousSubmission996/CLARITY upon acceptance. Certain production-specific services remain under NDA and cannot be distributed.
- **Lack of unified quantitative benchmarks.** There is still no publicly available dataset or protocol that simultaneously covers *triage*, *speciality selection*, *emergency detection*, and *safety filtering*. Numerical head-to-head evaluation with earlier systems is therefore impossible without re-implementing each approach on private clinical data. As a partial remedy we performed a *qualitative, functional comparison* of CLARITY with seven representative systems (AMIE, DISC-MedLLM, CoD, Polaris, MedAgents, Dual-Inf, and other). The resulting matrix - Appendix A

Table 3 - contrasts core capabilities and highlights CLARITY's unique FSM-based architecture. While this analysis cannot replace a common benchmark, it offers readers a transparent, like-for-like view of where our system stands relative to prior art and exposes gaps that future community datasets should address.

Next step. We will publish an open evaluation *protocol* (labeling guidelines, task definitions, scoring scripts) that lets any third party apply our four-task suite to *their own* data without sharing protected records. This avoids legal constraints on commercial medical logs while enabling reproducible, apples-to-apples benchmarking across independent institutions.

8 Ethics

Regulatory compliance. Under the applicable national data-protection legislation, studies that use *fully de-identified* service logs collected with *explicit informed consent* are classified as non-interventional quality-improvement research and therefore **exempt from mandatory Institutional Review Board review**. All analyses were conducted on anonymised records stored in an encrypted, access-controlled environment.

Informed consent and data handling. All users of the telehealth service explicitly agreed that their anonymised dialogue records may be used for research and quality-improvement purposes. Personally identifiable information (names, free-text IDs, contact details, timestamps finer than one hour) is removed at ingestion by an automated de-identification pipeline. Only the resulting anonymised logs were accessible to the research team.

Model safety and audit. CLARITY enforces a three-layer safety framework: (i) a domain-specific moderation filter, (ii) conservative emergency-detection thresholds, and (iii) deterministic FSM transitions that bound interaction depth. All prompts and model outputs are logged and periodically audited by licensed physicians.

9 Acknowledgments

We are grateful to **SberHealth** for supporting the project and to all participating physicians for their invaluable feedback during development and evaluation. We also thank our clinical partners and annotators for their assistance with data labeling, which was essential for validating and improving the study's outcomes.

References

Ali Ala, Vladimir Simic, Muhammet Deveci, and Dragan Pamucar. 2023. Simulation-based analysis of appointment scheduling system in healthcare services: A critical review. *Archives of Computational Methods in Engineering*, 30(3):1961–1978.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *Preprint*, arXiv:2308.14346.

G O Barnett, J J Cimino, J A Hupp, and E P Hoffer. 1987. DXplain. an evolving diagnostic decision-support system. *JAMA*, 258(1):67–74.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R Chaurasia, Nirav R Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A Pfeffer, and Nigam H Shah. 2024. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024. Cod, towards an interpretable medical agent using chain of diagnosis. *Preprint*, arXiv:2407.13301.

Martin J. Downes, Merehau C. Mervin, Joshua M. Byrnes, and Paul A. Scuffham. 2015. Telemedicine for general practice: a systematic review protocol. *Systematic Reviews*, 4(1):134.

Diwakar Gupta and Brian Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40:800–819.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

S Kühnel, M. Jovanović, H. Hoffman, S. Golde, L. Schneider, and M. Hirsch. 2023. Introduction of a pathophysiology-based diagnostic decision support system (DDSS) and its potential impact on the use of AI in healthcare.

Natalia Loukachevitch, Suresh Manandhar, Elina Baral, Igor Rozhkov, Pavel Braslavski, Vladimir Ivanov, Tatiana Batura, and Elena Tutubalina. 2023. NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities. *Bioinformatics*. Btad161.

Natalia Loukachevitch, Andrey Sakhovskiy, and Elena Tutubalina. 2024. Biomedical concept normalization over nested entities with partial umls terminology in russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2383–2389.

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, Jun Gao, Yuan-Geng-Shuo Wang, Jia-Ming Ji, Zifeng Qiu, Muzi Li, Cheng Qian, Tianze Guo, Shuangquan Ma, Zeying Wang, and 6 others. 2024. The application of large language models in medicine: A scoping review. *iScience*, 27(5):109713.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov, and Ehud Reiter. 2021. A preliminary study on evaluating consultation notes with post-editing. *Preprint*, arXiv:2104.04402.

Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, Cezanne Bianco, Swapnil Sharma, Rae Lasko, Michelle Voisard, Sanchay Harneja, Darya Filippova, Gerry Meixiong, Kevin Cha, Amir Youssefi, and 7 others. 2024. Polaris: A safety-focused llm constellation architecture for healthcare. *Preprint*, arXiv:2403.13313.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Aleksandr Nesterov, Andrey Sakhovskiy, Ivan Sviridov, Airat Valiev, Vladimir Makharev, Petr Anokhin, Galina Zubkova, and Elena Tutubalina. 2025. Ruccod: Towards automated icd coding in russian.

Alexandr Nesterov, Galina Zubkova, Zulfat Miftahutdinov, Vladimir Kokh, Elena Tutubalina, Artem Shelmanov, Anton Alekseev, Manvel Avetisian, Andrey Chertok, and Sergey Nikolenko. 2022. RuCCoN: Clinical concept normalization in Russian. pages 239–245.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *Preprint*, arXiv:2307.15343.

Xiaoming Shi, Zeming Liu, Li Du, and 1 others. 2024. Medical dialogue: A survey of categories, methods, evaluation and challenges.

E H Shortliffe. 1977. Mycin: A Knowledge-Based computer program applied to infectious diseases. *Proc Annu Symp Comput Appl Med Care*, pages 66–69.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. *Preprint*, arXiv:2311.10537.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Satvik Tripathi, Rithvik Sukumaran, and Tessa S Cook. 2024. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *J. Am. Med. Inform. Assoc.*, 31(6):1436–1440.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, and 6 others. 2024. Towards conversational diagnostic ai. *Preprint*, arXiv:2401.05654.

Dmitriy Umerenkov, Alexandr Nesterov, Vladimir Shaposhnikov, Ruslan Abramov, Nikolay Romanenko, Vladimir Kokh, Marina Kirina, Anton Abrosimov, Dmitry Dylov, and Ivan Oseledets. 2025. Ai diagnostic assistant (aida): A predictive model for diagnoses from health records in clinical decision support systems. pages 9880–9889.

Morgan Waller and Chad Stotler. 2018. Telemedicine: a primer. *Current Allergy and Asthma Reports*, 18(10):54.

Yifan Yang, Qiao Jin, Qingqing Zhu, Zhizheng Wang, Francisco Erramuspe Álvarez, Nicholas Wan, Benjamin Hou, and Zhiyong Lu. 2024. Beyond multiple-choice accuracy: Real-world challenges of implementing large language models in healthcare. *Preprint*, arXiv:2410.18460.

Qinkai Yu, Mingyu Jin, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, and Yongfeng Zhang. 2025. Health-llm: Personalized retrieval-augmented disease prediction system. *Preprint*, arXiv:2402.00746.

Hongjian Zhou, Fenglin Liu, Boyang Gu, and 1 others. 2024a. A survey of large language models in medicine: Progress, application, and challenge.

Shuang Zhou, Mingquan Lin, Sirui Ding, Jiashuo Wang, Genevieve B. Melton, James Zou, and Rui Zhang. 2024b. Interpretable differential diagnosis with dual-inference large language models. *Preprint*, arXiv:2407.07330.

A Related work

To describe the current state of medical dialogue systems, we first examine their evolution, highlighting how they have progressed from rule-based systems to modern LLM-powered AI. Then, we identify key challenges that must be addressed to develop robust and clinically reliable AI-driven consultations.

A.1 Evolution of Medical Dialogue Systems

Medical dialogue systems have evolved through three major stages:

1. **Rule-Based Expert Systems** – Early models such as Mycin (Shortliffe, 1977) and DXplain(Barnett et al., 1987) used manually encoded medical knowledge and decision trees. While interpretable, they lacked adaptability and could not handle open-ended patient input.
2. **Task-Specific AI Models** – Machine learning-based systems such as Babylon Health (Moramarco et al., 2021) and Ada Health (Kühnel et al., 2023) introduced data-driven triage and symptom assessment, but they remained single-turn and failed to engage in dynamic diagnostic hypotheses transparency.
3. **LLM-Powered Conversational Agents** – Recent systems, such as AMIE (Tu et al., 2024), DISC-MedLLM (Bao et al., 2023), CoD (Chen et al., 2024) and MEDAGENTS (Tang et al., 2024), integrate structured multi-turn dialogue processing and medical reasoning, improving diagnostic hypotheses accuracy and interaction depth.

Despite these advancements, medical dialogue systems still struggle with fundamental issues related to structured reasoning, factual reliability, and real-time decision-making.

A.2 Challenges in Medical AI Conversations

Medical dialogue systems have made significant progress, but their deployment in real-world clinical settings remains challenging. Despite leveraging LLMs for more natural and interactive patient communication, these systems exhibit critical limitations that hinder their reliability and effectiveness. The primary challenges include structured reasoning deficiencies, factual inaccuracies, and failure to prioritize critical cases.

A.2.1 Insufficient Transparency in Multi-Turn Diagnostic Dialogue

One of the primary shortcomings of current medical dialogue systems is their inability to conduct structured, multi-turn reasoning. Many models generate fragmented, inconsistent, or overly simplistic responses, failing to follow a logical sequence similar to a physician's diagnostic hypotheses inquiry.

Several approaches have been proposed to address the challenges of structured diagnostic hypotheses reasoning in medical dialogue systems. Chain of Diagnosis (CoD) introduces a framework designed to mimic the stepwise reasoning of physicians, systematically guiding the diagnostic hypotheses process and reducing the risk of premature conclusions (Chen et al., 2024). Another notable system, AMIE, leverages a self-play training mechanism within a simulated environment, enabling the model to iteratively refine both its questioning strategies and decision-making capabilities across a variety of medical scenarios (Tu et al., 2024). Similarly, DISC-MedLLM enhances the depth of structured questioning by integrating medical knowledge graphs and training on real-world, multi-turn medical dialogues, ensuring that the system can navigate complex patient interactions with greater precision (Bao et al., 2023).

Despite these advancements, challenges persist. Models struggle with handling ambiguous patient responses, dynamically adjusting diagnostic hypotheses strategies, and effectively incorporating context across long multi-turn conversations.

A.2.2 Hallucinations and Misinformation

A significant limitation of LLM-based medical dialogue systems is their tendency to generate factually incorrect or misleading information, known as hallucinations. These inaccuracies can lead to diagnostic hypotheses errors and compromise patient safety.

Several strategies have been proposed to mitigate the challenges associated with medical dialogue systems. Health-LLM utilizes a retrieval-augmented generation (RAG) mechanism to ground AI-generated responses in trusted medical sources, ensuring real-time verification and reducing the risk of hallucinations (Yu et al., 2025). In practice, retrieval and grounding can be supported by domain corpora that link clinical text to controlled terminologies (e.g., UMLS/ICD), such as NEREL-BIO, RuCCoN, and RuCCoD (Loukachevitch et al., 2023, 2024; Nesterov et al., 2022, 2025). Similarly, MEDAGENTS implements a multi-agent framework in which AI-based "experts" collaboratively cross-validate each other's responses, thereby minimizing incorrect outputs

and enhancing overall reliability (Tang et al., 2024). Another approach, DISC-MedLLM, combines structured multi-turn dialogues with the integration of medical knowledge graphs to provide detailed explanations for each diagnostic step. This design not only improves the transparency of the reasoning process but also facilitates validation by enabling clinicians to trace how conclusions are reached (Bao et al., 2023).

Nonetheless, hallucinations remain prevalent in complex diagnostic hypotheses cases, particularly those involving rare diseases, overlapping symptoms, or mental health conditions. Ensuring real-time fact-checking and clinician-in-the-loop validation remains a pressing challenge.

A.2.3 Failure to Recognize Critical Conditions

Medical AI systems often fail to detect and escalate high-risk conditions requiring urgent medical attention. This limitation reduces their effectiveness in emergency triage and high-stakes diagnostic hypotheses settings.

Several approaches have been developed to improve the recognition of critical conditions in medical dialogue systems. Dual-Inf enhances diagnostic hypotheses reliability by employing bidirectional inference, allowing the model to cross-reference symptoms with potential diagnostic hypotheses to detect severe conditions more accurately (Zhou et al., 2024b). Similarly, AMIE integrates an uncertainty-aware decision-making mechanism, which prompts the model to request additional clarifications whenever its confidence in a diagnostic hypotheses is low, thereby reducing the likelihood of incorrect or premature conclusions (Tu et al., 2024). Another approach, Polaris, leverages specialized risk assessment models to identify high-risk cases and escalate them for human intervention when necessary, ensuring that critical conditions receive appropriate attention (Mukherjee et al., 2024).

However, LLM-based medical systems still struggle with distinguishing time-sensitive symptoms from non-urgent ones, balancing AI autonomy with human intervention, and responding efficiently to emergency cases.

B Methods

B.1 Proposed Solution to Challenges

Our proposed system, CLARITY, offers a comprehensive solution to the challenges faced by existing medical dialogue systems, as summarized in Table 3. By integrating Finite State Machines (FSMs) with Large Language Models (LLMs), CLARITY outperforms other approaches across key dimensions: structured reasoning, safety, critical condition recognition, real-time readiness, adaptability, and scalability.

Transparency: CLARITY ensures consistent and structured multi-turn diagnostic hypotheses reasoning through FSMs, mimicking the systematic approach of experienced clinicians. In contrast, while systems like DISC-MedLLM and AMIE exhibit strong reasoning capabilities, they lack the robustness of FSM-driven workflows, leading to fragmented or inconsistent interactions in complex cases.

Safety: One of the most pressing challenges in medical AI is hallucination and misinformation. CLARITY addresses this through a Retrieval-Augmented Generation (RAG) approach, grounding all outputs in trusted medical knowledge bases. Systems like Polaris and Dual-Inf attempt similar mitigation strategies but fall short of the transparency and reliability offered by CLARITY.

Critical Condition Recognition: The probabilistic risk triage model in CLARITY ensures real-time identification and escalation of high-risk cases, outperforming solutions like MEDAGENTS, which lack the precision and adaptability needed for critical diagnostic hypotheses.

Real-time Readiness: CLARITY is designed for seamless integration into clinical environments, leveraging modular microservices for real-time deployment. Task-specific systems such as Babylon Health and Ada Health, while functional, are limited to specific use cases and lack the flexibility of CLARITY.

Adaptability: Unlike other systems, CLARITY dynamically adjusts to diverse scenarios, including rare diseases and ambiguous patient inputs, due to its hybrid architecture. LLM-based systems like Health-LLM and AMIE struggle in these areas due to limited context-awareness and reliance on static models.

Scalability: The modular design of CLARITY enables effortless scaling across various clinical infrastructures, making it suitable for both large hospital networks and smaller healthcare providers. Existing solutions, including DISC-MedLLM and Polaris, lack this level of deployment flexibility.

In summary, CLARITY demonstrates unmatched performance across all key criteria, bridging the gap between cutting-edge AI research and

Category	System	Transparency	Safety	Critical	Real-time	Adaptability	Scalability
Rule-based	MyCin(Shortliffe, 1977)	+-	+	+-	+	-	-
	Dxplain (Barnett et al., 1987)	+-	+	+-	+	-	-
Task-specified	Babylon Health (Moramarco et al., 2021)	-	-	+-	+	-	-
	Ada Health (Kühnel et al., 2023)	-	-	+-	+	-	-
LLM-based	DISC-MedLLM (Bao et al., 2023)	+	+-	+	+	+-	+-
	MEDAGENTS (Tang et al., 2024)	+-	+	+	+-	+-	-
	AMIE (Tu et al., 2024)	+	+	+	+-	+-	+-
	Health-LLM (Yu et al., 2025)	+-	+-	+	+-	-	-
	Polaris (Mukherjee et al., 2024)	+	+	+	+	+-	+-
	Dual-Inf (Zhou et al., 2024b)	+	+	+	+	+-	-
	CLARITY (**ours**)	+	+	+	+	+	+

Table 3: Qualitative capability matrix. Each '+' corresponds to a documented, peer-reviewed evaluation; '±' indicates partial support or missing evidence.

practical healthcare applications. By combining structured reasoning, advanced safety measures, and real-time adaptability, CLARITY sets a new standard for medical dialogue systems.

C Dialogue examples

D Detailed Component-Level Results

D.1 Moderator

The moderator was evaluated on a dataset of 10,000 messages, achieving an F_1-score of 0.95. The system demonstrated high accuracy in detecting prohibited topics, with a false positive rate of less than 1.5%.

D.2 Emergency Detector

We trained the emergency detection module on 2,114 medical dialogues and evaluated it on a held-out set of 682 chats, each independently annotated by three licensed physicians. At the operating threshold t=0.11, selected in alignment with stakeholder requirements for high reliability, the system achieved a precision of 0.72, an F_1-score of 0.59, and a false positive rate (FPR) of just 0.02. This threshold was deliberately optimized to minimize false alarms in real-world deployment, ensuring the system maintains a high level of trust and stability in routine clinical use. While the current configuration favors specificity, the model architecture allows for straightforward adjustment of this threshold to support more recall-oriented use cases, such as emergency triage scenarios. This adaptability makes the module suitable for diverse clinical workflows, and its consistent performance across evaluation settings provides a strong foundation for safe and effective integration.

Beyond the primary emergency detector, the safety stack includes:

- A domain-specific moderator that filters unsafe topics at input and output stages.
- Deterministic FSM gates that prevent transition to the "Diagnostic Hypothesis & Routing" state unless readiness is verified, holding low-confidence dialogues for human handoff.
- A universal red-flag questionnaire presented to all users as a final safety net.

D.3 Readiness Estimator

The readiness estimator was trained on a corpus of 2,500 medical dialogues, with each message labeled as part of the anamnesis collection or diagnostic hypotheses stage. The model used TF-IDF and contextual embeddings for feature extraction, achieving an F_1-score of 0.78. The mean absolute percentage error (MAPE) for predicting the duration of the anamnesis collection phase was 22%.

D.4 Question Detector

A total of 9,077 user messages were annotated to train the question detector. The classifier, based on logistic regression and contextual embeddings, achieved a recall of 0.87 for the positive class, a precision of 0.99 for the negative class, and a macro-averaged F_1-score of 0.94. These results demonstrate high accuracy in identifying clarifying questions.

D.5 Information Collector

The information collector module was tested on 1,000 dialogues, achieving a precision of 0.84 for question relevance. The average response time was 5 seconds, with 20% of cases utilizing cached questions from the historical dialogue database.

D.6 Medical Specialty Selector

Medical specialty selection system input $\mathfrak{W} = \{W_1, ..., W_n\}, n \in \mathbb{N}$ is a set of n medical documents W. W is a text string that contains a chat between a patient and a doctor. Medical specialty selection system output $\mathfrak{D} = [\![D_1, ..., D_n]\!], n \in \mathbb{N}$ is a multiset of n multisets $D = [\![d_1, ..., d_k]\!], k \in \mathbb{N}$,

	Transparency	Critical	Safety	Adaptability
Dialogue example	**Doc:** What's bothering you? **Pt:** I have a headache. **Doc:** Where exactly is the pain located? **Pt:** The back of my head. **Doc:** Are you experiencing any other symptoms, such as nausea, vomiting? **Pt:** No. **Doc:** How intense is the pain? **Pt:** 5 out of 10. **Doc:** Cervicogenic headache – General practitioner. Pain in the back of the head may be related to issues in the cervical spine. Cervical osteochondrosis – Neurologist. You have a neck condition that could be causing the headache in the back of your head. **Doc:** Is everything clear? Feel free to ask questions!	**Doc:** What is bothering you? **Pt:** High blood pressure. **Doc:** Do you experience chest pain or an increased heart rate? **Pt:** Yes. **Doc:** Do you have any vision problems? **Pt:** Yes. **Doc:** Your condition could be close to critical! Call 103 immediately. Wait for the response, do not hang up! Briefly explain what happened.	**Doc:** What's bothering you? **Pt:** Who are you? **Doc:** I am a neural network model that can assist with medical questions. How can I help you? **Pt:** Help me with algorithms in Python. **Doc:** I'm sorry, I didn't understand your response. Could you please provide more specific information or rephrase your message? **Pt:** What can you help me with? **Doc:** I can provide consultations on health and medical issues. Please tell me more about your symptoms or concerns so that I can offer appropriate assistance.	**Doc:** What's bothering you? **Pt:** I've been smoking for 5 years, half a pack a day. I've started coughing up very thick phlegm. How can I check my lungs? Could you provide a list of examinations and doctors I need to see? **Doc:** You need to see a general practitioner (therapist) or a pulmonologist first. The doctor will order a chest X-ray or a CT scan of the lungs to rule out pathological changes.
Related services metrics	Information collector: $P = 84\,\%$, $f_{\text{rep}} = 0$ Readiness estimator: MAPE $= 22\,\%$ Medical-specialty selector: $R@3 = 96\,\%$	Emergency detector: $F_1 = 59\,\%$, FPR $= 2\,\%$	Moderator: $F_1 = 95\,\%$	Question detector: $F_1 = 94\,\%$

Table 4: CLARITY dialogue examples and full-study results with relevant metrics.

where d is a text string that contains a doctor specialty. Expectations of system performance will be shaped by the set of experts $\mathfrak{E} = \{E_1, ..., E_z\}, z \in \mathbb{N}$, where expert is a function $E : \mathfrak{W} \mapsto \mathfrak{D}$. The system itself is a function $A : \mathfrak{W} \mapsto \mathfrak{D}$ which is not in a set of experts: $\chi_{\mathfrak{E}}(A) = \emptyset$. Then we can define the pairwise precision P and recall R between the algorithm and the expert as

$$P_{AE}@k = \frac{\sum_{i=1}^{n} \mu(D_i^A, D_i^E)}{nk^2} \tag{2}$$

$$R_{AE}@k = \frac{\sum_{i=1}^{n} \chi(D_i^A, D_i^E)}{n} \tag{3}$$

where $D_i^A = A(W_i)$, $D_i^E = E(W_i)$, μ is a multiplicity function, and χ is a characteristic function. We consider the accuracy of the model's answer by the complete match of its answer and the expert's answer. Thus, we can calculate the multiplicity μ and the characteristic χ functions in case of $|D^A| = |D^E| = k$:

$$\mu(D^A, D^E) = \sum_{p=1}^{k} \sum_{q=1}^{k} \begin{cases} 1 & d_p^A = d_q^E \\ 0 & d_p^A \neq d_q^E \end{cases} \tag{4}$$

$$\chi(D^A, D^E) = \begin{cases} 1 & \mu(D^A, D^E) > 0 \\ 0 & \mu(D^A, D^E) = 0 \end{cases} \tag{5}$$

The raw dataset is a collection of $n = 360$ chats. The first half of the dataset consists of chats from two doctors played out a pre-determined scenario. One of them knew the disease and their symptoms and was the patient. The other tried to define the disease by asking questions as the doctor. The second half of the dataset consists of the chats between the actor playing the role of a patient and CLARITY as the doctor. Ground truth labels for

the dataset were marked by $z = 7$ experts. Two of them are licensed doctors and other 8 are resident physicians. Each expert marked up all chats with up to $k_{max} = 5$ answers. The quality metrics presented in Table 5 are more than sufficient for product use. It is also interesting that the model ranks the answers in order of importance very well. The share of the first answer among the total number of correct answers is 80%, the share of the first two answers is 95%.

Table 5: Pairwise quality of the medical specialty selector compared to all experts. The number of specialists in the answer is equal to k.

k	Precision@k	Recall@k
1	77±0.3%	77±0.3%
2	71±0.2%	92±0.4%
3	68±0.3%	96±0.4%

D.7 Language Model Details

All LLM-based components (answer generation, diagnostic hypothesis drafting, and explanation synthesis) are powered by GigaChat. We used the production configuration provided by the vendor; prompts and safety constraints are listed in Appendix D. Where relevant, we apply deterministic decoding for safety-critical paths and temperature-limited sampling for open-dialogue responses.

E Formal Description of the FSM

Let the Finite State Machine be defined as a tuple $M = (Q, \Sigma, \Omega, C, T, \delta, \lambda, q_0, q_d, q_{ca}, F)$ where:

- Q is a finite set of states.
- W is the input alphabet, defined as the set of all possible strings representing user messages and previous chat history.
- Ω is the output alphabet, defined as the set of all possible text messages generated by the system.
- C is the set of all conditions for transitions.
- T is the set of all transitions.
- $\delta : Q \times W \to Q$ is the transition function.
- $\lambda : Q \times W \to \Omega$ is the output function.
- $q_0 \in Q$ is the initial state.
- $q_d \in Q$ is the default state.
- $q_{ca} \in Q$ is the state with the correct answer.
- $F \subseteq Q$ is the set of final states.
- $N\_ATTEMPTS$ is the maximum number of attempts allowed in the internal cycle.

E.1 Set of Conditions C

The set of conditions C is defined as:

$$C = \{c_1, c_2, \ldots, c_m\}$$

where each c_i is a condition that checks if the input $w \in W$ satisfies certain criteria.

E.2 Set of Transitions T

The set of transitions T is defined as:

$$T = \{t_1, t_2, \ldots, t_n\}$$

where each t_j is a transition defined by:

$$t_j = (q_j, c_j, q'_j)$$

where:

- $q_j \in Q$ is the source state.
- $c_j \subseteq C$ is the set of conditions for the transition.
- $q'_j \in Q$ is the target state.

Each transition t_j has an associated function $is\_triggered(w)$ that determines if the transition is triggered by the input $w \in W$. This function is defined as:

$$t_j.is\_triggered(w) = \bigwedge_{i=1}^{|c_j|} c_i(w)$$

E.3 Transition Function

The transition function δ is defined as:

$$\delta(q, w) = \begin{cases} q' & \text{if } \exists\, t \in T : \\ & \quad t.\,\text{is\_triggered}(w) = \text{true} \\ q_d & \text{otherwise} \end{cases}$$

where:

- T is the set of all transitions.
- $t.is\_triggered(w)$ checks if the input $w \in W^*$ received from the user satisfies the conditions of transition t.

E.4 Output Function

The output function λ is defined as:

$$\lambda : Q \times W \to \Omega$$

$$\lambda(q, w) = action_q(w)$$

where $action_q$ is the function associated with state q that generates the system response based on the input $w \in W^*$ received from the user.

Algorithm 1 FSM Algorithm

$q \leftarrow q_0$ ▷ Set the initial state
while not user_interrupted **do** ▷ External Cycle
  $w \leftarrow user\_input$ ▷ Get the user input
  $inner\_states \leftarrow \emptyset$ ▷ Initialize the inner states list
  while $q \notin F \wedge |inner\_states| < N\_ATTEMPTS \wedge q \neq q_{ca}$ **do** ▷ Internal Cycle
    $q' \leftarrow \delta(q, w)$ ▷ Apply the transition function
    $r \leftarrow \lambda(q', w)$ ▷ Generate the output
    $inner\_states \leftarrow inner\_states \cup \{(q', r)\}$ ▷ Add the state-output pair to inner states
    $q \leftarrow q'$ ▷ Update the current state
  end while
  return r ▷ Return the system response
end while

E.5 External Cycle

The external cycle represents the system-user interaction sequence:

$$(q_i, w_i) \vdash (q_{i+1}, w_{i+1}, r_i)$$

where:

- $q_i \in Q$ is the current state.
- $w_i \in W^*$ is the input string received from the user.
- $r_i \in \Omega$ is the system response.

The external cycle continues until the user interrupts it.

E.6 Internal Cycle

The internal cycle is defined as:

$$(q, w) \vdash^* (q', r)$$

where:

- $\vdash^*$ is the reflexive transitive closure of $\vdash$.
- q' is the state reached after at most $N\_ATTEMPTS$ steps.
- $r \in \Omega$ is the system response.

The internal cycle represents a single step in the external cycle, where the system generates a response based on the user input.

E.7 Termination Conditions

The internal cycle terminates when one of the following conditions is met:

$$q \in F \vee |inner\_states| = N\_ATTEMPTS \vee q = q_{ca}$$

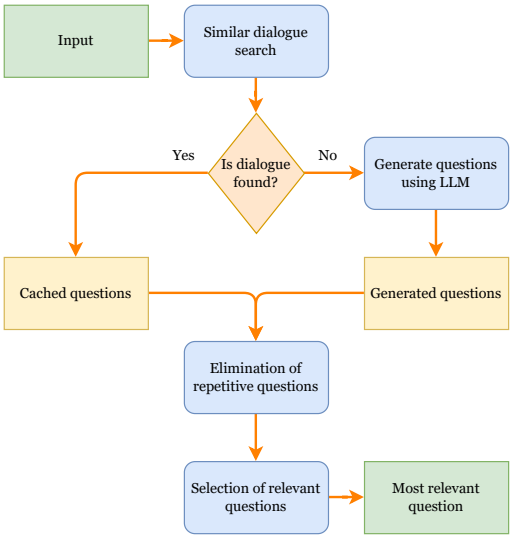
E.8 Output

The output of the internal cycle is defined as a tuple:

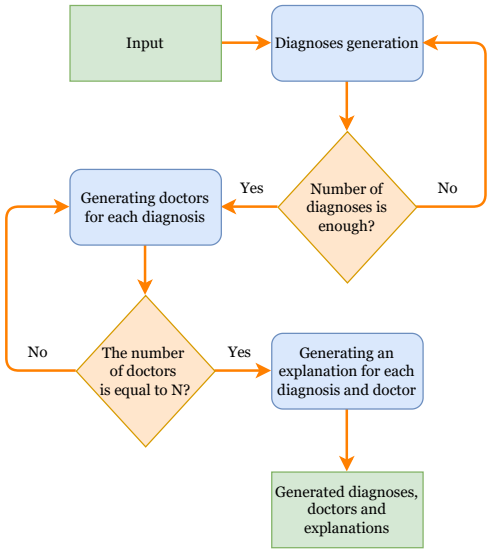
$$(r, q) \in \Omega \times Q$$

where r is the system response and q is the next state for the external cycle.

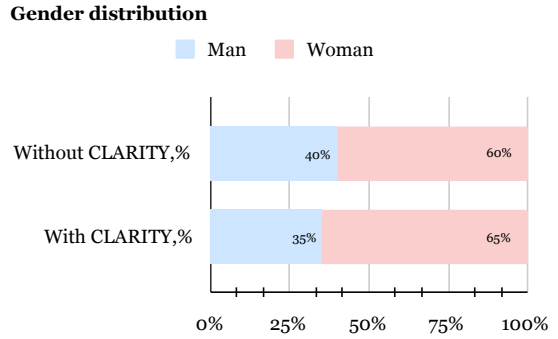
F Information Collector Diagram



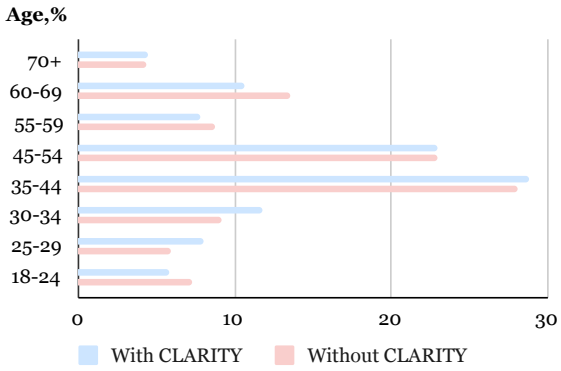
G Medical Specialty Selector Diagram



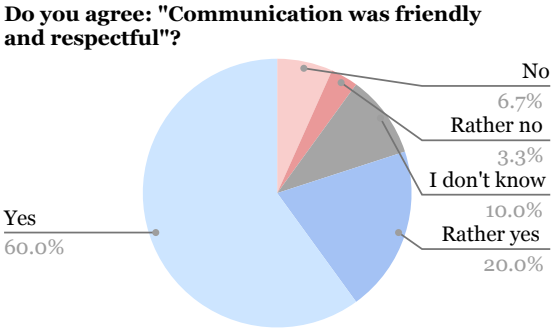
H Gender distribution



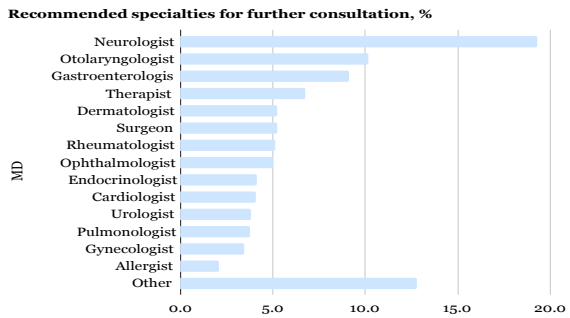
I Age groups distribution



J User survey results: interaction experience.



K Statistics on the recommended specialists in the expert-validated set of dialogues.



L User input description

The user interface of the CLARITY system exposes the `/v3/request` request URL for queries of various medical scenarios.

The user request body has the following contents

```
{
  "Text": "User message",
  "OuterContext": {
    "Sex": true,
    "Age": 21,
    "UserId": "UserId",
    "SessionId": "SessionId",
    "ClientId": "ClientId",
  }
},
```

where the `"Text"` field contains the input user message and the `"OuterContext"` field contains all the necessary information about the user. Inside the `"OuterContext"` the `"Sex"` and `"Age"` fields denote the user's age and sex, respectively. The `"UserId"`, `"SessionId"` and `"ClientId"` fields together form a unique dialogue identifier, which is used across the CLARITY components for dialogue context propagation.

M System output description

The CLARITY output format:

```
{
  "Text": "System output message",
  "Results": [...]
}
```

The `"Text"` field contains the system message. The `"Results"` field contains dialogue context and is empty except in the *Diagnostic hypotheses state*, where it contains three elements with this structure:

```
{
  "Diagnosis": "One of three diagnoses",
  "Doctor": "Recommended med specialty",
  "Description": "Diagnosis description"
},
```

where the `"Diagnosis"` field contains the name of the diagnosis, the `"Doctor"` field contains the suggested medical specialty for further consultation and the `"Description"` field contains text which clarifies the diagnosis name.