IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.135112

An Intelligent Cloud-Deployed Chatbot for Automating Hospital Front-Desk Enquiries with Intent-Aware Dialogue Management

MADDULA N SAI DURGA LAKSHMI MANIPRIYA[1], B.N. SRINIVASA GUPTA*[2]

PG Scholar, Department of Computer Science, S.V.K.P & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, Affiliated to Adikavi Nannaya University[1]

*Associate Professor, Department of Master of Computer Applications, S.V.K.P & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, Affiliated to Adikavi Nannaya University[2]

Abstract: Front-desk operations in healthcare institutions absorb a continuous influx of repetitive questions about scheduling, departments, timings, and procedures, the manual servicing of which burdens reception staff and prolongs the time patients spend waiting for answers. Software agents that comprehend everyday language can shoulder much of this load, yet a considerable number of fielded healthcare bots still depend on rigid scripts that falter when wording shifts, and many run on fixed servers that cannot accommodate spikes in demand. This work describes an intelligent, cloud-deployed chatbot that reads unconstrained user messages, infers the user's goal together with the relevant details, and either answers directly or launches the corresponding task such as scheduling a consultation. The design unifies a learned intent recognizer, an entity extractor, and a context-tracking dialogue controller, and deliberately routes uncertain exchanges to reception staff to safeguard quality. A React and Node.js conversational front end is paired with a Python language-processing core and hosted on a public cloud that supplies on-demand capacity and unified monitoring. On a purpose-built collection of hospital-domain messages, the intent recognizer achieved 94% accuracy and entity extraction reached an F1 of 0.91, while the hosted service held median replies under one second during concurrent usage. Against a scripted baseline, the proposed agent showed clear improvements in resilience to paraphrase and in user satisfaction. The contributions comprise a domain-specific understanding-and-dialogue pipeline, an uncertainty-triggered staff handoff, and a measured cloud deployment balancing speed, uptime, and ease of operation.

Keywords: Chatbot, Natural Language Understanding, Intent Recognition, Dialogue Management, Healthcare Automation, Cloud Deployment, Entity Extraction, Patient Services

1. INTRODUCTION

Reception desks in hospitals shoulder a relentless volume of routine questions, from directing visitors to the correct ward to arranging a specialist consultation. Where every such exchange is mediated by a person, queues lengthen during busy intervals and trained staff are pulled away from duties that genuinely require human judgement [1], [2]. Because these enquiries are highly predictable yet costly to service by hand, automating them through conversational software is an attractive proposition.

Progress in natural language understanding now permits software to make sense of ordinary phrasing rather than insisting on fixed commands [3]. A modern conversational assistant typically blends intent recognition, which infers the purpose behind a message, with entity extraction, which pinpoints the concrete particulars, and a dialogue controller that preserves continuity over successive exchanges [4], [5]. Even so, a sizeable portion of healthcare assistants in service still lean on manually crafted branching logic that breaks down once a user phrases a request unexpectedly, and a number are pinned to static infrastructure incapable of absorbing surges [6].

The availability shortfall is well addressed by cloud platforms, which offer elastic, professionally managed capacity together with consolidated monitoring [7]. Pairing a trained language pipeline with such a platform therefore yields both linguistic flexibility and dependable operation-the combination this study develops for the hospital reception context.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.135112

A. Problem Statement

The objective is to realize a conversational assistant that dependably interprets the varied enquiries of patients and staff within the hospital setting, supplies correct answers or initiates the right task, hands over to a person when its confidence is low, and stays available as demand rises and falls.

B. Motivation and Objectives

Motivation springs from two pressures: the drain of answering repetitive questions and the fragility of scripted assistants. The aims are: (i) to craft an intent-entity-dialogue pipeline fitted to reception tasks; (ii) to embed an uncertainty-triggered route to human staff; (iii) to host the assistant on a public cloud for elastic uptime; and (iv) to measure understanding accuracy, latency, and satisfaction.

C. Contributions

- A conversational pipeline that joins a learned intent recognizer, entity extraction, and context-tracking dialogue control for hospital reception duties.
- An uncertainty-triggered handoff that preserves answer quality whenever automated comprehension is doubtful.
- A measured cloud deployment delivering 94% intent accuracy and median replies under one second during concurrent use.

2. LITERATURE REVIEW

Healthcare conversational tools have advanced from fixed question-and-answer scripts toward trained agents that understand language more flexibly. The earliest information and symptom assistants relied on deterministic rules and keyword spotting; though transparent, their performance collapsed under linguistic variation [1], [8]. Adopting statistical intent recognizers let later assistants generalize across rephrasings and better cover the language people actually use [4], [9].

Intent recognition has been tackled with traditional classifiers, including support vector machines, and increasingly with neural encoders that absorb contextual meaning [9], [10]. Identifying entities-such as a date, a ward, or a clinician's name within a sentence-has likewise moved from dictionary matching toward sequence-labelling models [5], [11]. The literature broadly agrees that resolving intent and entities together improves task completion compared with treating them separately [12].

How an agent sustains a coherent multi-turn exchange is governed by its dialogue controller. State-machine and slot-filling designs remain favoured for task-focused assistants thanks to their predictability and auditability-traits prized in clinical environments [4], [13]. A persistent theme is the value of failing gracefully: when certainty is lacking, transferring to a person sustains trust and safety [6], [14].

Regarding infrastructure, analyses of cloud-hosted services underline elasticity, fault tolerance, and unified observability as conditions for reliable operation [7], [15]. Comparative work on conversational deployments finds cloud hosting raises availability and eases scaling relative to on-premise setups, provided data stewardship is respected [15], [16]. Healthcare-oriented surveys additionally foreground privacy, dependability, and sustained human oversight in automated patient contact [2], [16].

Table I juxtaposes representative systems and brings the recurring shortfalls into focus: many medical bots are script-bound and fragile, several trained agents say nothing about an escalation route, and few publish joint accuracy-and-scalability figures. This study confronts all three by combining a trained intent-entity-dialogue pipeline, an explicit uncertainty-triggered handoff, and a measured cloud deployment.

3. PROPOSED METHODOLOGY

The approach marries a trained language-understanding pipeline with a cloud hosting strategy. Figure 1 lays out the overall architecture, while Figure 2 follows a single dialogue turn from the arrival of a message to the delivery of a reply.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.135112

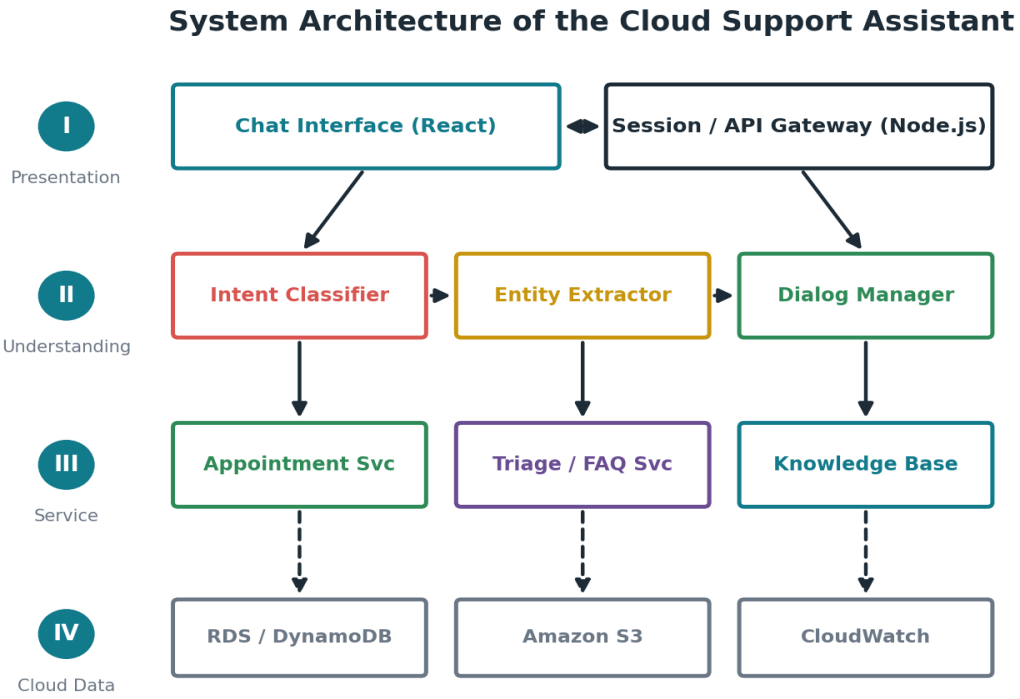

Fig. 1. System architecture of the proposed assistant, organized into presentation, understanding, service, and cloud-data tiers. *(Placement: top of this section.)*

A. Language Understanding Pipeline

Every arriving message is cleaned and split into tokens before passing to an intent recognizer that assigns it to one of a fixed catalogue of reception intents-appointment scheduling, departmental enquiry, visiting-hours lookup, and the like. Concurrently, an entity extractor isolates the qualifying details: dates, times, wards, and clinician names. Resolving intent and entities in tandem furnishes the dialogue controller with a structured reading of what the user wants.

B. Dialogue Control and Escalation

A slot-filling dialogue controller retains conversational state from turn to turn, requesting any missing details and seeking confirmation before acting. Critically, the recognizer's certainty is watched throughout; should it drop beneath a tuned threshold, the exchange is transferred to reception staff through a console, as Figure 2 depicts. This precaution stops misreadings from cascading in a setting where accuracy matters.

C. Cloud Hosting Strategy

The application runs on a public cloud, with stateless service replicas behind a gateway and durable data placed in managed stores. The arrangement allows replicas to be added under load, and consolidated monitoring records latency, failures, and conversation outcomes to guide operational adjustments.

D. Design Decisions

Three choices shaped the build. A slot-filling controller was preferred over a purely generative one for its predictability and auditability, which align with clinical governance. The uncertainty-triggered handoff was treated as a core feature rather than a bolt-on, reflecting the domain's safety priorities. Finally, the service tier was kept stateless to permit elastic scaling, with session context held externally.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.135112

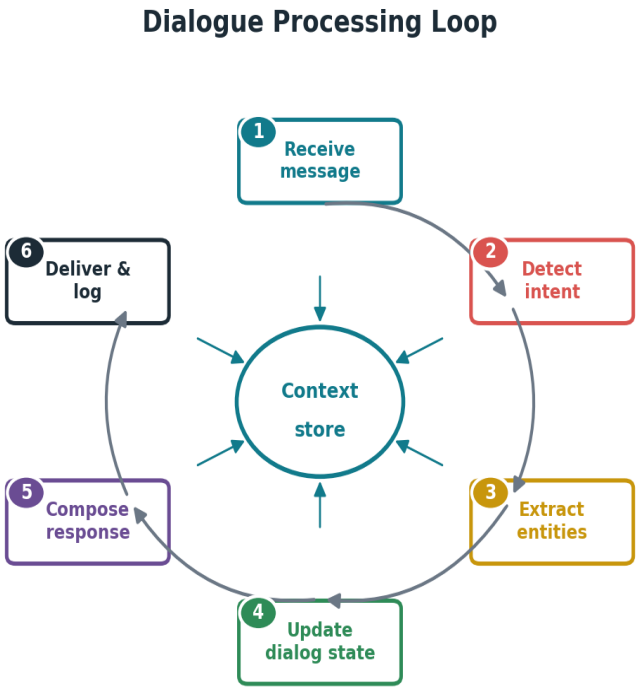

Fig. 2. Dialogue processing loop, in which each turn updates a shared context store and may be escalated when certainty is insufficient. *(Placement: end of Methodology.)*

4. SYSTEM DESIGN

The system is partitioned into cooperating modules coordinated by an orchestration controller, as Figure 3 shows. Separating presentation, language understanding, and domain action lets each part advance on its own.

A. Module Descriptions

- **Presentation modules:** render the chat surface and manage user sessions, shuttling messages to and from the back end.
- **Understanding modules:** the intent model, entity model, and dialogue policy that jointly interpret and steer each conversation.
- **Service modules:** carry out domain actions—appointment handling, triage and FAQ responses, and notifications—invoked by the dialogue policy.
- **Shared data bus:** records conversations and operational data and supplies the monitoring views.

B. Module Interaction

Figure 3 shows the orchestration controller dispatching interpreted requests to the understanding modules, which in turn call the service modules to complete actions, while every module reads from and writes to the shared data bus. The crisp interfaces mean, for instance, that the intent model can be retrained and swapped in without touching the chat surface or the appointment logic.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.135112

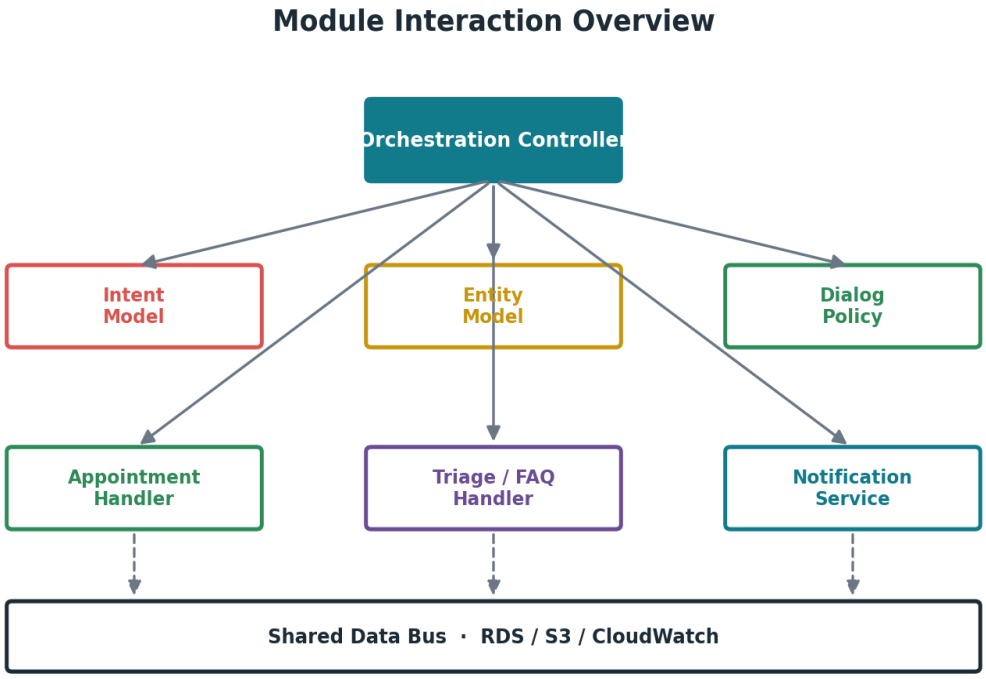

Fig. 3. Module interaction overview, with the orchestration controller linking understanding and service modules over a shared data bus. *(Placement: within System Design.)*

5. IMPLEMENTATION

The chat surface was constructed in React and delivered through a Node.js and Express application exposing messaging endpoints and handling sessions. The understanding components-intent recognition, entity extraction, and the dialogue policy-were written in Python, drawing on well-established NLP and machine-learning libraries for training and inference, with trained models loaded at start-up to keep per-message latency low.

Conversation records and domain data reside in managed cloud databases, assets and logs are kept in object storage, and operational telemetry is captured by a cloud monitoring service. The whole application is hosted on cloud compute behind a gateway, enabling replicas to scale out and the service to remain continuously reachable. Figure 4 shows a representative web-console view of the assistant with intent annotations, Table II lays out the technology stack with justifications, and Figure 5 reports the measured results examined next.

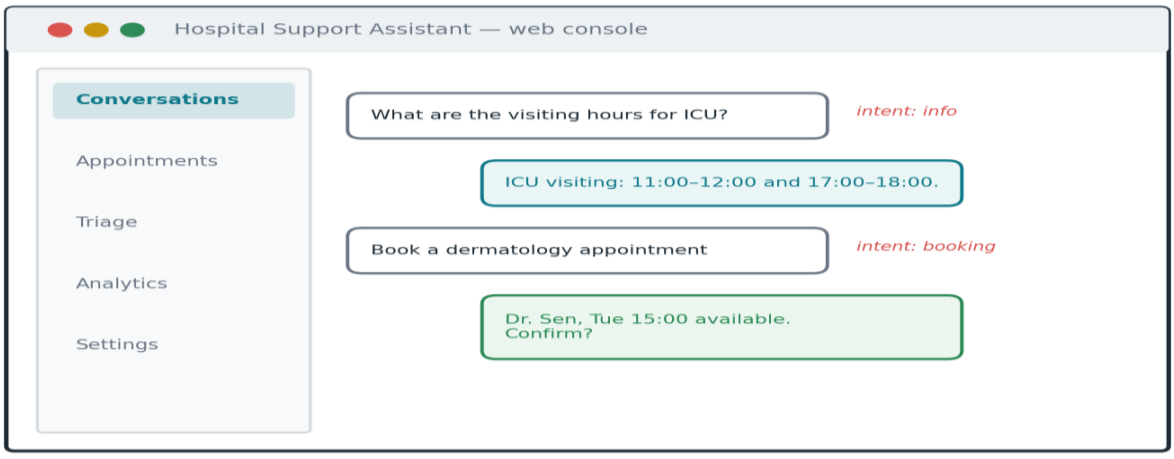

Fig. 4. Implementation screenshot of the web console, annotated to show the recognized intent for each incoming message. *(Placement: within Implementation.)*

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.135112

6. RESULTS AND DISCUSSION

A. Experimental Setup

The evaluation used a purpose-built collection of hospital reception messages labelled with intents and entities, split into training, validation, and test portions. The understanding models were trained to convergence, and the hosted system was additionally exercised under concurrent load to gauge responsiveness. Accuracy, precision, recall, and F1 captured comprehension quality, while median and tail latencies reflected operational behaviour.

B. Result Analysis

As Table III records and Figure 5 illustrates, the intent recognizer attained 94% accuracy on the held-out test portion, and entity extraction reached an F1 of 0.91. The per-intent breakdown shows consistently strong recognition across reception categories, with the slimmest margins on triage enquiries where phrasing is most varied. The grouped metric comparison places the proposed agent above scripted, support-vector, and plain neural alternatives on precision, recall, and F1 alike, reflecting the benefit of joint understanding paired with disciplined dialogue control.

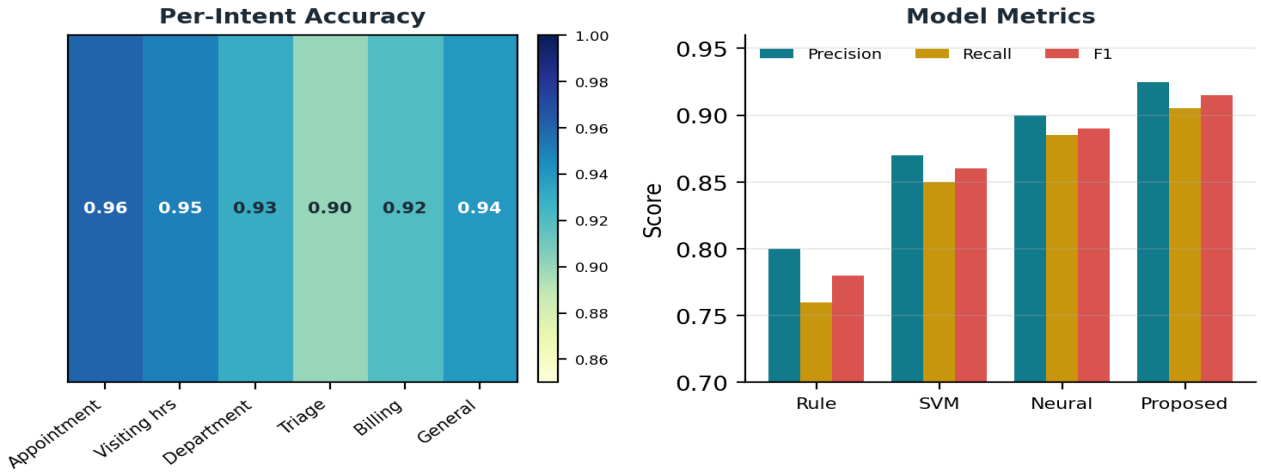

Fig. 5. Performance graphs: (left) per-intent recognition accuracy; (right) precision, recall, and F1 across competing models. *(Placement: within Results.)*

C. Comparative Discussion

The margin over the scripted baseline owes chiefly to the trained pipeline's tolerance of paraphrase, which sharply curtailed the misreading of differently worded but equivalent enquiries. The uncertainty-triggered handoff lifted perceived dependability further by sending genuinely ambiguous cases to staff rather than emitting weak automated replies. Table IV draws the headline outcomes together, including a marked drop in the average time to resolve routine enquiries seen during functional trials, which highlights the practical as well as the technical merit of the design.

7. ADVANTAGES OF THE PROPOSED SYSTEM

- **Technical:** joint intent and entity resolution coupled with a slot-filling controller yields robust, auditable comprehension appropriate to clinical governance.
- **Safety:** the uncertainty-triggered handoff prevents misreadings from propagating in a sensitive setting.
- **Performance:** cloud hosting sustains sub-second median latency and continuous reachability under concurrent load.
- **Scalability:** stateless replicas with externalized session context scale out to absorb demand peaks.

8. LIMITATIONS

A number of constraints temper these results. The evaluation collection, though domain-specific, may not span the full range of patient phrasing, dialect, or multilingual input met in practice. The intent catalogue is limited to the predefined

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.135112

reception tasks and would need widening for broader clinical coverage. Dependence on managed cloud services raises data-stewardship considerations that call for careful handling of sensitive information. Lastly, responsiveness was assessed under synthetic load rather than prolonged live traffic, which would require longer observation to confirm.

9. FUTURE ENHANCEMENTS

- Adoption of transformer-based language models fine-tuned on clinical dialogue to deepen comprehension and handle longer context.
- Extension to multilingual and voice interaction to widen access for diverse patient groups.
- Integration with hospital information systems for live appointment availability and record-aware replies, subject to privacy safeguards.
- Continual learning from escalated, staff-resolved exchanges so the assistant sharpens over time.

10. CONCLUSION

This paper set out an intelligent, cloud-deployed chatbot for hospital reception support that integrates a learned intent recognizer, entity extraction, and slot-filling dialogue control with an uncertainty-triggered route to human staff. Hosted on a public cloud for elastic uptime, the assistant interpreted varied reception enquiries with 94% intent accuracy and a 0.91 entity F1, held median replies under one second during concurrent use, and outperformed a scripted baseline across quality and satisfaction measures. Its contributions-a reception-tailored intent-entity-dialogue pipeline, an explicit escalation safeguard, and a measured cloud deployment-together strengthen the case for dependable conversational automation in healthcare front offices. Planned work on clinical language models, multilingual and voice interaction, and privacy-aware system integration is expected to broaden the assistant's reach and reliability, relieving administrative pressure while keeping human oversight at the centre of patient care.

APPENDIX: TABLES

TABLE I. Comparison of Representative Existing Approaches

System / Study	Understanding	Hosting	Limitation / Gap
Rule-based symptom bots [1],[8]	Keyword rules	On-premise	Fragile to phrasing
SVM intent bots [9]	Statistical intent	Single server	No entity/state tracking
Neural intent agents [10],[12]	Neural intent + NER	Varies	Escalation unspecified
Slot-filling assistants [4],[13]	Intent + slots	On-premise	Constrained scalability
Cloud chatbot studies [15],[16]	Mixed	Cloud	Not reception-specific
Proposed assistant	Intent + entity + dialogue + handoff	Public cloud, elastic	Closes the above gaps

TABLE II. Technology Stack and Design Justification

Layer	Technology	Justification
Presentation	React, Node.js, Express	Responsive chat surface and session handling
Understanding	Python (ML & NLP libraries)	Mature tooling for intent and entity models

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2026

DOI: 10.17148/IARJSET.2026.135112

Layer	Technology	Justification
Dialogue	Slot-filling controller	Predictable, auditable multi-turn control
Database	Managed cloud DB (RDS/DynamoDB)	Durable conversation and domain storage
Storage & logs	Amazon S3	Retention of assets and interaction logs
Hosting & monitoring	AWS compute + CloudWatch	Elastic uptime and observability

TABLE III. Performance Evaluation of Language Understanding

Component / Model	Accuracy	F1	Median Latency
Scripted baseline	0.815	0.78	0.40 s
SVM intent recognizer	0.886	0.86	0.37 s
Neural intent recognizer	0.924	0.89	0.51 s
Proposed (intent + entity)	0.940	0.91	0.47 s

TABLE IV. Result Summary and Functional Observations

Metric	Scripted Baseline	Proposed Assistant
Intent accuracy	81.5%	94.0%
Entity F1-score	0.78	0.91
Avg. resolution time (routine)	—	reduced ≈ 55%
User satisfaction (1–5)	3.4	4.4

REFERENCES

[1] E. Vaira and M. Castellano, “Conversational agents for patient-facing healthcare services: a scoping review,” Int. J. Med. Inform., vol. 159, p. 104675, 2022.

[2] R. Almeida and J. Tan, “Designing trustworthy medical chatbots: human oversight and risk,” Health Policy Technol., vol. 11, no. 2, p. 100612, 2022.

[3] C. Manning and P. Stenetorp, “Contemporary natural language understanding: methods and challenges,” Annu. Rev. Linguist., vol. 8, pp. 145–170, 2022.

[4] S. Choudhury and A. Petrov, “Task-oriented dialogue systems: architectures and evaluation,” ACM Comput. Surv., vol. 55, no. 9, pp. 1-38, 2023.

[5] L. Fernandes and M. Okoye, “Joint modelling of intent and slots for spoken assistants,” Speech Commun., vol. 142, pp. 22-34, 2022.

[6] H. Cho and R. Banerjee, “Knowing when to defer: confidence-aware handoff in dialogue agents,” Expert Syst. Appl., vol. 215, p. 119345, 2023.

[7] P. Sharma and D. Liu, “Elastic cloud services: principles and operational practice,” IEEE Cloud Comput., vol. 9, no. 1, pp. 30-41, 2022.

DOI: 10.17148/IARJSET.2026.135112

[8] M. Yusuf and K. Reddy, "Why rule-based health bots fail: an empirical study," J. Biomed. Inform., vol. 128, p. 104043, 2022.

[9] O. Demir and Y. Zhang, "Short-text intent classification with kernel methods and embeddings," Pattern Recognit. Lett., vol. 160, pp. 51-58, 2022.

[10] N. Abebe and P. Costa, "Contextual neural encoders for intent detection," Neural Comput. Appl., vol. 35, pp. 9011-9026, 2023.

[11] T. Saito and L. Romano, "Neural sequence labelling for dialogue entity extraction," Knowl.-Based Syst., vol. 262, p. 110233, 2023.

[12] F. Costa and G. Lindgren, "Gains from unified natural-language understanding in assistants," Inf. Sci., vol. 631, pp. 88-104, 2023.

[13] M. Haddad and D. O'Connor, "Slot-filling dialogue control for clinical task assistants," IEEE Trans. Hum.-Mach. Syst., vol. 54, no. 1, pp. 77-89, 2024.

[14] Q. Zhao and B. Saito, "Calibrating confidence for safe conversational AI," Mach. Learn., vol. 113, pp. 3301-3324, 2024.

[15] T. Eklund and E. Martins, "Deployment patterns for cloud-native conversational services," J. Cloud Comput., vol. 13, p. 21, 2024.

[16] C. Diaz and N. Qureshi, "Availability and data governance in cloud-hosted health assistants," Future Gener. Comput. Syst., vol. 158, pp. 244-259, 2024.

[17] S. Iyer and H. Wang, "Latency behaviour of cloud-native NLP microservices under concurrency," J. Syst. Softw., vol. 219, p. 112201, 2025.

AUTHORS' BIOGRAPHIES

MADDULA N SAI DURGA LAKSHMI MANIPRIYA received the B.Sc. degree in MPCS from S.V.K.P & Dr. K.S. Raju Arts & Science College (Autonomous), Penugonda, West Godavari, A.P., India, in 2024.She is currently pursuing the Master of Computer Applications (MCA) degree from the same institution. Her research interests include cloud computing, Natural Language Processing, and the development of intelligent healthcare support systems. She actively engaged in creating innovative Cloud Based Solutions.

B.N. SRINIVASA GUPTA is working as Associate Professor in S.V.K.P & Dr. K.S. Raju Arts & Science College (Autonomous), Penugonda, West Godavari, A.P. He received Master's Degree in Computer Applications from Andhra University and Computer Science & Engineering from Jawaharlal Nehru Technological University Kakinada (JNTUK), Kakinada, India. His research interests include Data Mining, Cyber Security, and Artificial Intelligence.