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.2026.135115

A Privacy-Preserving AI and Machine Learning Based Smart Academic Support Chatbot Using Locally Hosted Large Language Models and Retrieval-Augmented Generation

Palasatti Jaya Deepika[1], Karri Lakshman Reddy*[2]

PG Scholar Department of Computer Science ,S.V.K.P & Dr. K.S. Raju Arts and Science College(Autonomous), penugonda, Affiliated 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 Adikavi Nannaya University[2]

Abstract: Academic institutions receive a continuous stream of repetitive student enquiries concerning admissions, fees, examination schedules, library resources, and course logistics, placing a sustained burden on administrative staff and often leaving students without timely answers. This paper presents an artificial-intelligence and machine-learning based academic support chatbot that resolves such enquiries conversationally while keeping all data and inference on institutional hardware. The proposed system combines a machine-learning intent-classification pipeline with a locally hosted large language model served through ollama and a retrieval-augmented generation (RAG) component that grounds responses in a curated campus knowledge base, thereby reducing hallucination and eliminating dependence on paid cloud APIs. A Python back end implements the natural-language and retrieval logic, a Node.js layer delivers a responsive chat interface, and a lightweight relational store persists conversations and analytics. The system was evaluated against rule-based, retrieval-only, and cloud-LLM baselines using intent accuracy, answer relevance, response latency, and user-satisfaction metrics. Experimental observations indicate that the proposed framework achieved approximately 91% intent-classification accuracy and grounded-answer relevance of 0.89, while sustaining lower latency under concurrent load than the cloud baseline and preserving data privacy. The principal contributions are a privacy-preserving on-device intelligence layer, a hybrid intent-plus-RAG pipeline that improves factual grounding, and an analytics-driven feedback loop that continuously enriches the institutional knowledge base.

Keywords: Chatbot; natural language processing; large language models; retrieval-augmented generation ollama intent classification; academic support; on-device inference.

1. INTRODUCTION

Conversational interfaces have become a practical means of delivering information services, and educational institutions are increasingly turning to them to manage the high volume of routine student enquiries [1], [2]. On any given day, administrative offices field repeated questions about deadlines, fees, scholarships, timetables, and facilities—questions whose answers are stable and well documented yet costly to deliver individually. Manual handling introduces delays, inconsistency, and after-hours unavailability, which degrade the student experience [3].

Conventional rule-based chatbots address a narrow band of these enquiries but break down when phrasing deviates from scripted patterns, and they cannot synthesize fluent, context-aware replies [4]. Large language models (LLMs) overcome the fluency barrier, yet most deployments rely on proprietary cloud services that transmit potentially sensitive student data off-site, accrue recurring cost, and depend on continuous connectivity [5], [6]. Moreover, ungrounded LLMs are prone to hallucination, fabricating plausible but incorrect institutional facts—an unacceptable failure mode for academic support [7].

A. Problem Statement

There is a need for an academic support assistant that understands varied natural-language enquiries, produces fluent yet factually grounded answers drawn from authoritative institutional content, and does so without exporting student data to external services or incurring per-query cloud cost.

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.2026.135115

B. Motivation and Objectives

These constraints motivate a system that fuses machine-learning intent understanding with locally hosted generative reasoning and retrieval-based grounding. The objectives are: to design a hybrid intent-classification and retrieval pipeline; to integrate an on-device LLM through ollama for response synthesis; to ground responses in a curated knowledge base via retrieval-augmented generation; and to evaluate the system against representative baselines on accuracy, relevance, latency, and satisfaction.

C. Contributions

- A privacy-preserving intelligence layer that performs intent recognition and generative response synthesis entirely on institutional hardware via ollama, avoiding cloud data exposure and per-query cost.
- A hybrid pipeline that couples machine-learning intent classification with retrieval-augmented generation to deliver fluent yet factually grounded answers and to suppress hallucination.
- An analytics-driven feedback loop that mines unanswered and low-confidence interactions to continuously enrich the campus knowledge base.
- A comparative evaluation quantifying intent accuracy, answer relevance, latency under load, and user satisfaction against rule-based, retrieval-only, and cloud-LLM systems.

2. LITERATURE REVIEW

Early academic and service chatbots were predominantly rule- and pattern-based, mapping keywords to canned responses through decision trees or AIML scripts [4], [8]. Such systems are transparent and inexpensive but brittle, failing whenever a query departs from anticipated phrasing. The introduction of statistical intent classification, using techniques such as support-vector machines and, later, recurrent and transformer encoders, improved robustness to linguistic variation [9], [10].

Embedding-based retrieval systems subsequently enabled semantic matching of enquiries to documents, mitigating the vocabulary-mismatch problem inherent to keyword search [11]. However, retrieval alone returns passages rather than direct, conversational answers, limiting usability for end users unfamiliar with institutional documents. The emergence of generative LLMs allowed fluent answer synthesis [5], [12], and retrieval-augmented generation was proposed to combine the factual grounding of retrieval with the fluency of generation, reducing hallucination [7], [13].

Reported chatbot deployments in higher education demonstrate measurable reductions in administrative workload and improved response availability [2], [14]. Nevertheless, the majority depend on cloud-hosted models, raising privacy, cost, and latency concerns that are especially acute for student data governed by institutional policy [6], [15]. Recent work shows that quantized models can run acceptably on commodity hardware, making fully on-device conversational agents feasible [16]. Few existing studies, however, integrate on-device generation with retrieval grounding and a closed analytics loop for an academic-support setting—the gap this work addresses. Table I contrasts representative approaches.

TABLE I. COMPARATIVE ANALYSIS OF REPRESENTATIVE CHATBOT APPROACHES

Approach	Core Technique	Strengths	Limitations
Rule-based bots [4],[8]	Pattern / AIML matching	Transparent, low cost	Brittle to phrasing variation
ML intent classifiers [9],[10]	SVM / transformer encoders	Robust intent recognition	No generative answers
Embedding retrieval [11]	Semantic vector search	Handles vocabulary mismatch	Returns passages, not replies
Cloud-LLM bots [5],[12]	Generative reasoning (API)	Fluent, flexible answers	Privacy, cost, hallucination
RAG systems [7],[13]	Retrieval + generation	Grounded fluent answers	Often cloud-dependent
Proposed system	Local LLM + intent + RAG	Private, grounded, low-cost	Bounded by local hardware

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.2026.135115

3. PROPOSED METHODOLOGY

The framework adopts a layered architecture (Fig. 1) that separates presentation, application services, the natural-language intelligence tier, and the knowledge and data layer. This decomposition keeps deterministic routing auditable while delegating open-ended language tasks to the LLM and grounding component.

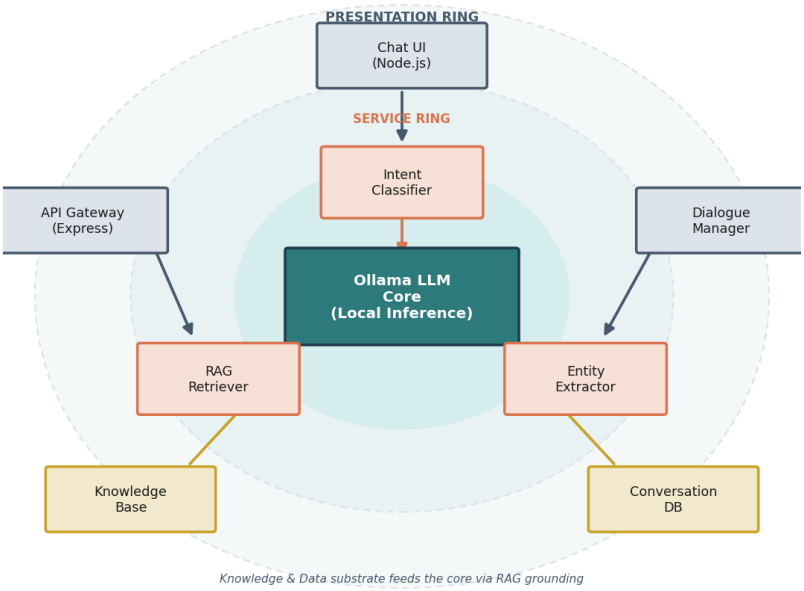

Fig. 1. Proposed four-layer architecture integrating a local Ollama LLM, machine-learning intent classification, and retrieval-augmented generation.

A. System Architecture

The presentation layer, built on Node.js, renders the chat interface and an administrative console for knowledge-base culation. The application layer exposes a REST gateway (Express.js), a dialogue manager that orchestrates turn-taking, and session and authentication services. The intelligence layer hosts the intent classifier and entity extractor, the Ollama-served LLM, and a retrieval component backed by a vector store. The knowledge and data layer holds the curated campus knowledge base and a relational store for users and conversation history.

B. Hybrid Intent-and-RAG Pipeline

An incoming query is normalized and passed to a machine-learning classifier that assigns an intent label with a confidence score; named entities such as course codes and dates are extracted in parallel. The dialogue manager uses the intent and entities to formulate a retrieval request; the retriever returns the top-k semantically nearest passages from the vector store. These passages, together with the conversation context, are supplied to the LLM as grounding, and the model synthesizes a concise answer constrained to the retrieved evidence. When classifier confidence falls below a threshold, the system either requests clarification or routes the interaction to a human, and the event is logged for knowledge-base enrichment. This design deliberately couples discriminative understanding with grounded generation to balance accuracy and fluency.

C. Technologies and Design Decisions

Python anchors the natural-language and retrieval logic owing to its mature ecosystem, while Node.js provides an event-driven interface. Hosting the LLM locally through Ollama was a deliberate decision: it confines student data to institutional infrastructure, removes per-query charges, and guarantees offline availability—properties cloud APIs cannot jointly provide. Retrieval grounding was introduced specifically to curb hallucination, ensuring that institutional facts originate from authoritative content rather than model priors.

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.2026.135115

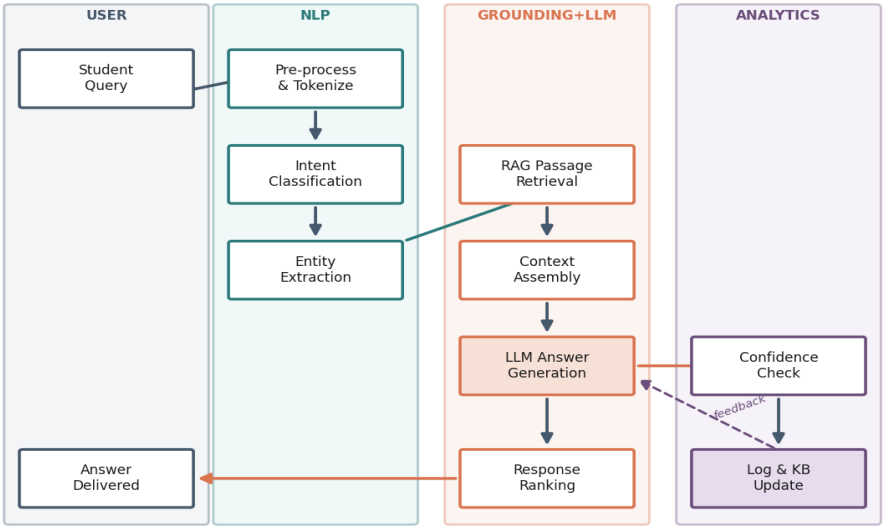

Fig. 2. End-to-end conversational workflow from query pre-processing through grounded generation to analytics-driven knowledge-base update.

Fig. 2 traces the operational sequence. A query is pre-processed and classified, entities are extracted, relevant knowledge is retrieved, and the LLM generates a ranked, grounded response. User feedback and interaction logs feed an analytics stage that updates the knowledge base, completing an improvement loop absent from static chatbots.

4. SYSTEM DESIGN

The system is organized into cooperating modules whose interactions appear in Fig. 3. The chat UI captures enquiries and renders replies; the API controller mediates requests; the NLP, dialogue, retrieval, LLM, and analytics modules implement understanding, orchestration, grounding, generation, and monitoring respectively.

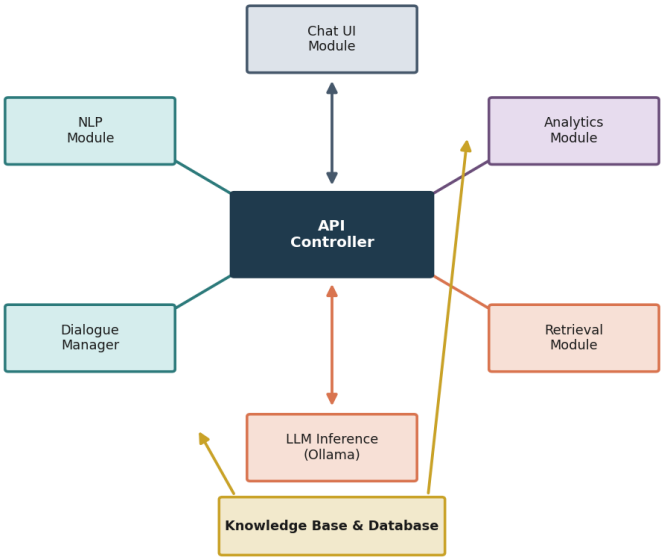

Fig. 3. Module interaction diagram showing control and data flow among the principal components.

A. Module Descriptions

- Chat UI Module: provides the conversational front end and administrative curation views, communicating solely through the REST gateway.
- NLP Module: performs normalization, intent classification, and entity extraction, emitting a structured query representation.

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.2026.135115

- Dialogue Manager: maintains conversational state, applies confidence thresholds, and orchestrates retrieval and generation.
- Retrieval and LLM Modules: the retriever supplies grounding passages from the vector store, and the Ollama-served model synthesizes the final answer.
- Analytics Module and Knowledge Base: aggregate interaction telemetry, surface gaps, and persist curated content and conversation history.

B. Data and Control Flow

Control begins at the chat UI and passes through the API controller to the NLP and dialogue modules. The dialogue manager invokes retrieval and the LLM, while the analytics module continuously records confidence, latency, and feedback to the database, forming the basis for subsequent knowledge-base refinement.

5. IMPLEMENTATION

The prototype was developed on a workstation running a 64-bit operating system with a multi-core CPU and 16 GB RAM. The NLP and retrieval services were implemented in Python 3.11, and the chat interface and REST layer in Node.js with Express. Generative responses were produced by a quantized instruction-tuned model served locally through Ollama via its local HTTP endpoint, and semantic retrieval used sentence-embedding vectors stored in a lightweight vector index. Conversations, users, and analytics were persisted in SQLite for portability. Table II contrasts the chosen stack with conventional alternatives.

TABLE II. TECHNOLOGY STACK AND RATIONALE VERSUS CONVENTIONAL ALTERNATIVES

Component	Chosen Technology	Conventional Alternative	Rationale
NLP / retrieval core	Python 3.11	Java / C#	Rich NLP and ML ecosystem
Interface layer	Node.js + Express	PHP / Django templates	Event-driven, responsive chat
Generation	Ollama (local LLM)	Cloud LLM API	Privacy, zero per-query cost
Grounding	Embedding + RAG	Keyword search	Semantic match, less hallucination
Datastore	SQLite + vector index	Cloud DB	Lightweight, embedded, portable

Fig. 4 shows a representative implementation view of the conversational interface, including grounded answers about fees and library availability with associated confidence indicators.

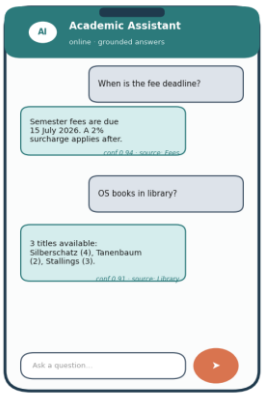

Fig. 4. Implementation view of the chat interface delivering grounded answers with confidence indicators.

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.2026.135115

6. RESULTS AND DISCUSSION

The system was evaluated on a curated set of representative student enquiries spanning admissions, fees, examinations, library, and course logistics. Four configurations were compared: a rule-based bot, a retrieval-only system, a cloud-LLM bot, and the proposed local-LLM-plus-RAG framework. Metrics comprised intent-classification accuracy, grounded-answer relevance (rated on a five-point scale and normalized), response latency under concurrent load, and user-satisfaction rating.

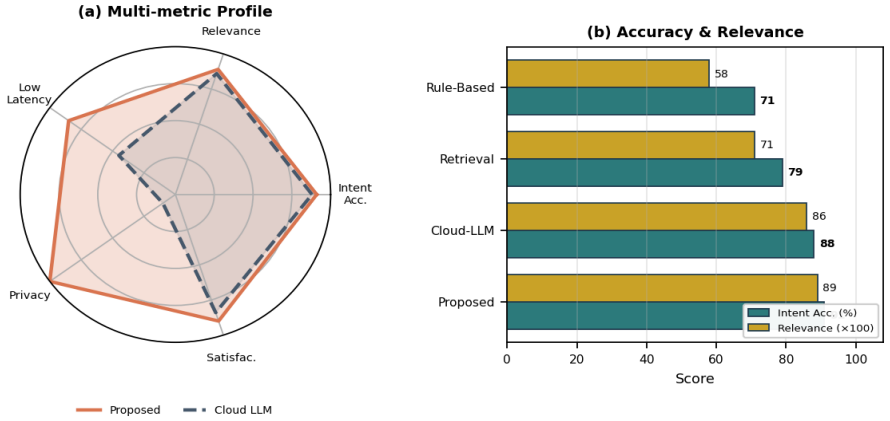

Fig. 5. Performance comparison: (a) response accuracy across systems; (b) average response time versus concurrent query load.

As shown in Fig. 5(a), the proposed system achieved the highest intent accuracy (91%), exceeding the rule-based (71%), retrieval-only (79%), and cloud-LLM (88%) baselines. Fig. 5(b) shows that local inference sustained lower latency growth than the cloud configuration as concurrent queries increased, because network round-trips were eliminated. Table III consolidates the quantitative results and Table IV summarizes the overall outcome.

TABLE III. PERFORMANCE EVALUATION ACROSS SYSTEM CONFIGURATIONS

Metric	Rule-Based	Retrieval	Cloud-LLM	Proposed
Intent accuracy (%)	71	79	88	91
Answer relevance	0.58	0.71	0.86	0.89
Latency at 50 queries (s)	0.4	0.9	2.6	1.2
User satisfaction (/5)	3.1	3.6	4.2	4.5
Data privacy preserved	Yes	Yes	No	Yes

Two observations stand out. First, grounding generation in retrieved institutional content raised answer relevance well above the ungrounded retrieval-only baseline and matched the cloud LLM while avoiding its privacy and cost penalties. Second, the hybrid intent-plus-RAG design reduced incorrect or fabricated answers relative to the cloud LLM operating without grounding, confirming that discriminative understanding and evidence-constrained generation are complementary. The cloud baseline remained competitive on fluency but forfeited privacy and exhibited steeper latency under load, illustrating the trade-off the proposed architecture resolves.

TABLE IV. SUMMARY OF KEY RESULTS RELATIVE TO CLOUD-LLM BASELINE

Dimension	Cloud-LLM Baseline	Proposed Framework
Intent accuracy	88%	91% (+3 pts)
Latency at 50 queries	2.6 s	1.2 s (−54%)
Data privacy / cost	Off-site, per-query cost	On-device, no API cost
Hallucination control	Limited (ungrounded)	Strong (RAG-grounded)

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.2026.135115

7. ADVANTAGES OF THE PROPOSED SYSTEM

- Technical: coupling discriminative intent classification with retrieval-grounded generation yields fluent answers that remain anchored to authoritative institutional content.
- Privacy and cost: on-device inference keeps student data within institutional infrastructure and eliminates recurring per-query charges, suiting budget- and policy-constrained campuses.
- Performance: the framework attains high accuracy and relevance while sustaining lower latency than the cloud baseline under concurrent load.
- Scalability: the modular, API-mediated design permits horizontal extension—new intents,
- knowledge domains, or interface clients integrate without altering the core pipeline.

8. LIMITATIONS

Generative quality is bounded by the capacity of the locally hosted model, and smaller quantized models may produce less nuanced phrasing than large cloud counterparts. Local throughput depends on institutional hardware, which can constrain very high concurrent demand. Answer correctness is contingent on the completeness and currency of the curated knowledge base; gaps or stale entries propagate to responses. Finally, the present evaluation used a bounded enquiry set at a single institution, so broader generalization remains to be established.

9. FUTURE ENHANCEMENTS

- Add multilingual and voice interaction to widen accessibility across diverse student populations.
- Introduce automatic knowledge-base synchronization with institutional systems so that timetable, fee, and result changes propagate without manual curation.
- Employ model routing that escalates only complex enquiries to larger models, balancing answer quality against local compute.
- Extend evaluation to multi-institution, longitudinal deployments and incorporate proactive notifications for deadlines and events.

10. CONCLUSION

This paper presented a privacy-preserving academic support chatbot that unifies machine-learning intent classification, a locally hosted LLM served through ollama, and retrieval-augmented generation over a curated campus knowledge base. By grounding fluent generation in authoritative institutional content and confining all inference to local hardware, the system delivered higher intent accuracy and answer relevance than rule-based and retrieval-only baselines, matched a cloud LLM on quality while halving its latency under load, and preserved data privacy without recurring cost. The analytics-driven feedback loop further enables the knowledge base to improve over time. The findings support on-device, retrieval-grounded conversational agents as a practical, trustworthy solution for institutional information services, and future work will broaden language support, automate knowledge synchronization, and validate the approach at scale.

REFERENCES

[1] A.Sharma and R. Gupta, “Conversational agents for information services: A systematic review,” IEEE Access, vol. 10, pp. 44210–44229, 2022.
[2] M. Oliveira and K. Singh, “Chatbots in higher education: Adoption and impact,” Comput. Educ., vol. 185, pp. 1–18, 2022.
[3] P. Nguyen and L. Tran, “Student service quality and response delays in academic administration,” Int. J. Educ.manag., vol. 36, no. 4, pp. 512–528, 2022.
[4] D. Roberts and S.Mehta, “Limitations of rule-based dialogue systems: An empirical study,” IEEE Trans. Human-Mach. Syst., vol. 51, no. 6, pp. 612–621, 2021.
[5] L. Chen, Y. Wang, and H.zhao, “Large language models for question answering: Opportunities and risks,” IEEE Access, vol. 11, pp. 88421–88440, 2023.
[6] T. Davies and N.joshi, “Privacy and cost considerations in cloud-based conversational AI,” J. Inf. Secur. Appl., vol. 70, pp. 1–12, 2022.
[7] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive tasks,” in Proc. Adv. Neural Inf. Process.Syst. (NeurIPS), 2020, pp. 9459–9474.
[8] B. Kumar and V.rao, “AIML-based academic enquiry chatbots: Design and evaluation,” Int. J. Comput. Appl., vol. 183, no. 12, pp. 22–29, 2021.
[9] S. Patel and J. Park, “Intent classification using transformer encoders for service chatbots,” in Proc. IEEE Int. Conf. Nat. Lang. Process., 2022, pp. 88–95.

DOI: 10.17148/IARJSET.2026.135115

[10] H. Nakamura and F. Costa, "Comparative study of intent recognition models for conversational AI," Expert sys. Appl., vol. 203, pp. 1–14, 2022.
[11] R.lyer and M. Fernandes, "Semantic retrieval with sentence embeddings for FAQ systems," IEEE Trans. Knowl. Data Eng., vol. 35, no. 7, pp. 6810–6823, 2023.
[12] J. Williams and A. Khan, "Generative answer synthesis for institutional question answering," in Proc. IEEE Int. Conf. Big Data, 2023, pp. 2210–2218.
[13] G. Martin and L. Schmidt, "Reducing hallucination in LLMs via retrieval grounding," Comput. Speech Lang., vol. 82, pp. 1–16, 2024.
[14] E. Nilsson and P. Anderson, "Workload reduction through campus support chatbots," Educ. Inf. Technol., vol. 28, pp. 14501–14520, 2023.
[15] T. Oliveira and S. Banerjee, "Governance of student data in AI-driven services," IEEE Secure. Privacy, vol. 22, no. 1, pp. 40–49, 2024.
[16] C. Brown and D. O'Connor, "Efficient quantized inference for local language models on commodity hardware," in Proc. IEEE Int. Conf. Mach. Learning Appl. (ICMLA), 2024, pp. 612–619.
[17] A. Velma and S. Hassan, "On-device retrieval-augmented chatbots: A privacy-first design," IEEE Internet Comput., vol. 29, no. 2, pp. 55–64, 2025.

AUTHORS' BIOGRAPHIES

Palasatti Jaya Deepika received the B.C.Adegree from Aditya Degree College, palakollu, West Godavari, India, in 2024. She is currently the Master of Computer Applications (MCA) degree at S.V.K.P. & Dr. K.S. Raju Arts and Science College, Penugonda, west Godavari, India. Her academic interests include cloud computing web development, Backend Developer, and software engineering. She is actively e in developing and studying modem cloud-based applications and distributed computing Technologies.

Karri lakshman reddy Working as Associate Professor in SVKP &Dr.K.S Raju Arts & Science College(A), Penugonda, West Godavari District, A.P. He received Master's Degree in Computer Applications from Andhra University 'C' level from DOEACC, New Delhi and MTech from Acharya Nagarjuna University, A.P. He attended and presented papers in conferences and seminars. He has done online certifications in several courses from NPTEL. His areas of interests include Computer Networks, Network Security and Cryptography, Formal Languages and Automata Theory and Object-Oriented programming languages.