Bridging Language Barriers in Education: A Literature Review on Multilingual and Code Mixed Chatbot Technologies

Addressing the Language Gap in Educational Help-based on AI Technology.

T. Harika, Kaniz-E-Fatima, Sofia Samreen

Assistant Professor, Student, Student
Artificial Intelligence & Data Science & Computer Engineering (ADCE),
Stanley College of Engineering and Technology for Women, Hyderabad, India
harikathalla@gmail.com, kanizfatima1504@gmail.com, sofiasam144@gmail.com

Abstract*— AI technological advancements combined with NLP advancements drove intelligent automated student support to replace digital higher education help services. Educational institutions use question generation tools extensively for repetitive information requests in areas such as placements and exams and educational funding and academic schedule information and student accommodation policies and student support services. Heavy academic situation demand creates delays which unites email helpdesk requests with offline help desks and phone helpline requests unable to support every student adequately. Most present-day conversational systems remain English-centric which means limited language capability for their users. Current chatbot technologies create incorrect responses in India's multilingual student communication because students speak code-mixed language but also use informal terms and transliterated writing styles. Throughout this literature review we study newer advancements toward multilingual and code-mixed conversational AI systems and present important methodologies like transformer-based language modelling with cross-lingual semantic representation for token-level language identification and Retrieval-Augmented Generation (RAG). The review presents the advantages and limitations of current multilingual educational chatbots and examines persistent research challenges including translation noise together with dialect handling and language switching inconsistencies and context retention issues and absence of document-backed reasoning for institutional reliability. The study finds that academic support systems in this field require combination of language processing components with human review modules and contextual memory-based storage for reliability and universal support. Research suggests the essential need for building language-neutral chatbots for students which support continuous conversation across regional languages and incorporation of code-mixed inputs between users. Under its active operation the system effectively expands educational access alongside streamlining administrative work for universities and delivers reliable data to higher-education institutions nationwide

***Index Terms*— Multilingual chatbots, Code-mixed, NLP, language-agnostic, AI, retrieval-augmented generator (RAG), Cross-lingual communication, Conversational AI for education, Human-in-the-loop support systems, Automation of Student assistance.**

I. INTRODUCTION

Higher education institutions operate as central learning bodies that direct students through both admission protocols as well as scholarship programs together with registration deadlines and university facilities and financial fees. Growing student numbers in colleges and universities create huge volumes of redundant help desk inquiries as well as repeated questions sent to faculty members and administrative staff and faculty staff via emails. The workforce at support systems based on traditional techniques fails to handle extensive client requests thus leading to client dissatisfaction coupled with poor service turnaround time. आंखों के जांच निर्देश करता है कि applications of this limitation cause maximum effect on especially students demographics throughout admission season registration period for exams and result issuing time span.

Artificial Intelligence (AI) systems together with Natural Language Processing (NLP) frameworks respond to student inquiries using chatbots or conversational agents to bring automatic resolution to these issues. These systems understand natural language user expressions to decrease administrative burdens by answering questions independently. Present educational chatbots use either English-only training data with rule-based user intent classification methods for query processing. Template-based or fixed keyword chatbots struggle to respond appropriately when users present questions in free-flowing natural language formats. The language environment of India complicates matters when students select native-language communication or English code-mixed expressions like Hindi-English (Hinglish) or Tamil-English (Tamlish) or Telugu-English. In their informal chats students extensively use code-mixed language yet these forms are expected to become their standard college-day discussion language. Standard architecture for multilingual chatbots cannot process code-switching so systems fail to identify user intent properly and produce incorrect response patterns. Translation approaches move meaning but eliminate crucial contextual and cultural elements found only in regional expressions. Large language models combined with transformer-based multilingual NLP models enable chatbots to exceed monolingual limitations during both understanding and response generation processes. Direct applications with LLMs in student support systems present implementation challenges that result from practical constraints. Educational chatbots have two main functions: identify code-mixed questions from multilingual students then deliver official document-based answers from school documents because inaccurate details severely impact students' academic outcomes. Retrieval-based architectures must be employed to access institutional validated data instead of complete reliance on generative language models. Academic chatbots need to achieve seamless continuity when conducting conversations. Students send continuous follow-up questions

related to their original messages. Most current chatbots handle each message separately which creates religious interruptions during conversations and weakens system performance. Established chatbot platforms do not feature any established protocols to route complex matters towards human professionals for case management. Software support agents must seamlessly redirect students to human support agents since either their answers lack sufficient confidence or the student's question involves private information. The educational industry requires an original document-based multilingual chatbot solution that recognizes code-mixing practices among native languages and supplies context-sensitive replies derived from specialized documents. The system should handle language differences to cut down on human-based support requests while supplying reliable information that boosts access and efficiency between students and administrative personnel. Language inclusivity determines access to academic support in countries with diverse linguistic populations so removing this gap is essential for these nations' academic support services.

II. BACKGROUND OF MULTILINGUAL CONVERSATIONAL AI

The conversational artificial intelligence technology enables machines to communicate with users by sending messages through English-language text and creating simulated human dialogs. The subsequent winding-down chatbot strategy combined simple keyword searches with predetermined pattern rules; natural language processing progress together with deep learning algorithms now allow chatbot platforms to analyze user contexts while including emotional tone and different ways of speaking. A rapid response function with exact educational data needs to be part of AI systems utilized for schools because their language diversity comprehension creates authentic support to maintain core educational success.

Development Timeline of NLP in BOT users

The modern process of transforming conversational AI into distinct stages consists of three different developmental phases:

Table 1: Timeline of chatbots

Generation	Core Technology	Capacity	Limitations
Gen 1	Rule-based chatbots	Simple keyword triggers & scripted replies	No learning, rigid responses
Gen 2	Machine learning + Intent classification	Classifies user intent & generates predefined responses	Requires large labelled datasets
Gen 3	Pretrained transformer-based models (BERT, GPT, mBERT)	Better language understanding & context awareness	Still struggles with domain-specific accuracy
Gen 4(Current)	Large Language Models + RAG retrieval	Natural conversation & document-grounded reasoning	Expensive models, domain adaptation required

Chatbots upgraded their response capabilities because they shifted from analyzing key words towards understanding deeper message meanings.

UNIVERSAL AI

Multilingual AI's core architecture has no need for training sessions of each language because this system can comprehend many languages. The creation of multilingual transformer models like mBERT along with XLM-R or mT5, LaBSE and LLaMA-2 multilingual multiple languages provides a common cross-lingual embedding space to enable:

- Languages zero-shot learning
- Linguistic common vocabulary related language pairs
- Language transfer learning from rich resource to scarce resource languages

The technological breakthrough enables chatbots to function optimally throughout the multiple languages which predominate in India's regions.

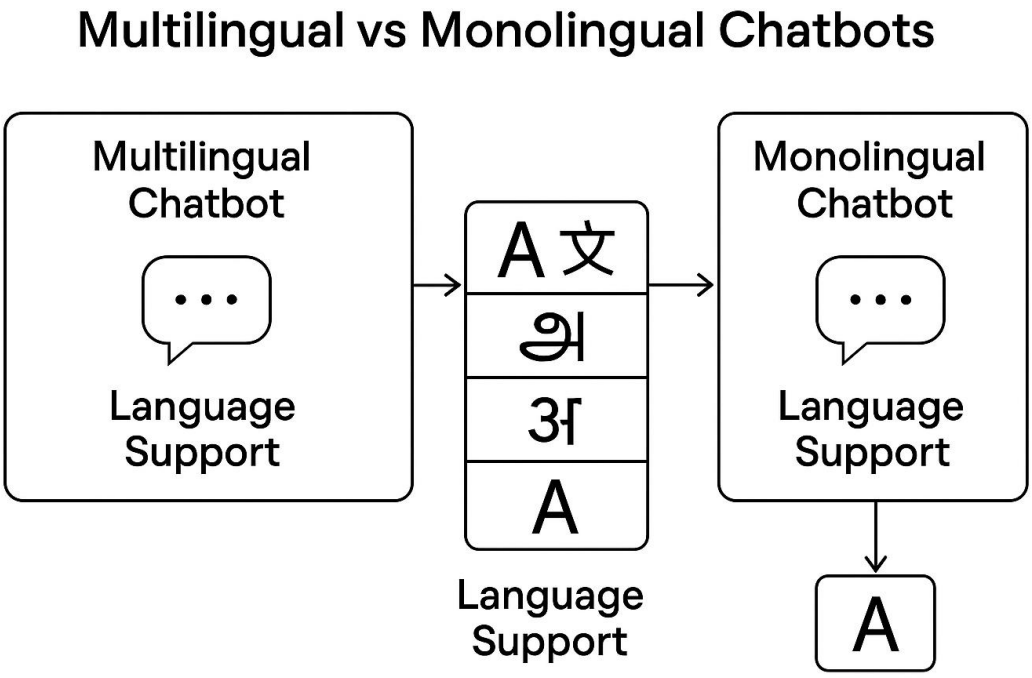

Fig. 1: Architecture of multilingual chatbot

HOW CODE-MIXED NEIGHBORHOOD COMMUNICATION WORKS

Code-mixing refers to a spontaneous exchange from different languages that occurs inside one sentence. Examples If you want to say:

- "Admission ke liye last date kya hai?"
- "Hostel fees kab submit karni hai?"
- “Scholarship form Tamil la fill panna help pannunga?”

The presence of this language pattern shows its prevalence among Indian students. Code-mixed texts also create several issues including:

- Economic sentence patterns
- Complete Hindi word writing/Letters manipulation from Hindi to English alphabets
- Unconventional spellings
- Language switching skyrocketed at mid-sentence

The inherent complexity keeps pure-language dataset-trained multilingual approaches and translation-based chatbots from recognizing meaning.

RETRIEVAL-AUGMENTED GENERATION FOR ACCURATE ANSWERS

Students require more than chatter they need accuracy. Through the following synergy Retrieval-Augmented Generation builds more reliable chatbots:

- Retrieval Model → Obtain most relevant text segment from academic institutional materials
- Generator (LLM) → Generate final response based on obtained documents

Hallucination prevention enables chatbot deliver only real academic information to users.

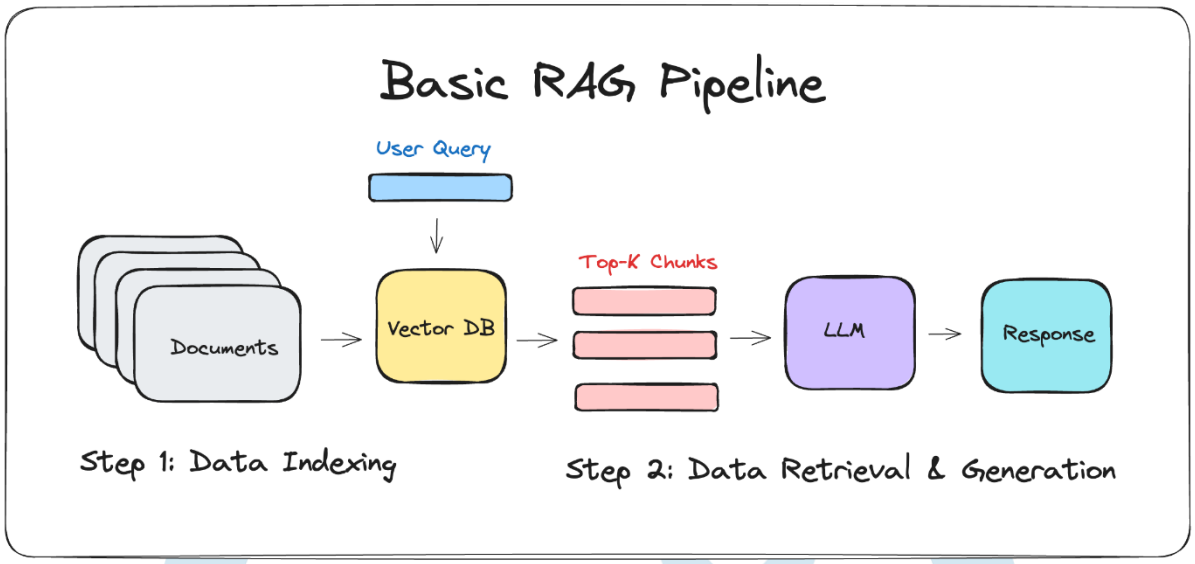

Fig. 2: RAG pipeline

CONTEXT-AWARE CONVERSATIONS

Students ask different follow-up questions to discover earlier information because their inquiry style differs from customer support chatbots that operate like this:

- When is application deadline?
- When can diploma students apply through admission deadline?

A resilience chatbot has to monitor information regarding:

- Get past conversation records
- Language used by users
- Previous question asked
- Session happening currently

The undisruptive maintenance of conversation continuity keeps communication moving smoothly.

HUMAN-IN-THE-LOOP CONVERSATIONAL AI

Chatbots developed through advanced language modeling experience the following problems:

- Substandard question construction
- Emotion-based queries
- Human inspection required questions

Human-in-the-loop systems redirect clients smoothly when confidence levels drop to maintain consistent quality service along with credible answers.

III. CHATBOT CAPABILITIES FOR UNDERPINNING EDUCATIONAL TEAMS

Educational chatbots operate as student-parent-staff virtual help lines for receiving quick informational and support access. Through FAQ answering assistance chatbots build a digital campus network which drives up engagement and user satisfaction and functional performance for educational institutions.
Chatbots in education handle student message troubleshooting by specializing in these types of explicit communication:

- Enrolment & Admissions: Dates for admission and entrance requirements and submitted papers and guidance program appointments
- Academic Services & Assessments: Details about schedules and curriculum components and evaluation methods and assessment results
- Financial & Scholarship Matters: Information about tuition and enrollment deadlines and institutional protocols and award application procedures
- Campus Operations & Facilities: Regulations about student housing and library services and campus transportation and student engagement activities and career placement resources

Automated systems allow academic staff time from recurring questions so they direct their work toward complex challenges.

HOW EDUCATION BENEFITS THE INSTITUTIONS

- 24×7 Availability: Provides accessibility outside office timing
- Reduced Workload: Staff get relief from continuous standard requests
- Consistency: Ensures delivery of uniform policy-based responses
- Scalability: Maintains workload capacity for admission/exam period peak traffic

- Analytics: Monitors frequent client questions to enhance institutional communication

MOVING STUDENTS TOWARD SUCCESS

- Quick responses beat typical email or office reply timelines
- User support for languages and code-mixing in content submissions
- Enabling students to seek further details without experience of fear
- Academic services resembling on-campus support offered remotely to students

EDUCATIONAL CHATS FUNCTIONAL REQUIREMENTS ANALYSIS

The proper operation of chatbots requires multiple elements.

- Links to Shipping Documents (brochures and circulars and regulations)
- Multi-lingual and code-mixed query handling
- Tracking conversation context through multiple discussion turns
- Sending program-specific responses as well as student-related responses
- Humans receive chats if automatic verification fails established thresholds
- Students' data security meets two regulatory requirements

CURRENT EDUCATIONAL CHATBOTS' DEFECTS

Modern platforms show these challenges:

- Design that focuses solely on English language content
- Default intent-based frameworks struggle to engage with non-direct questions
- Current systems do not emerge from document frameworks of educational institutions
- Blended user language inputs create roadblocks for computer processes

Fragile conversation structures stem from inadequate contextual awareness

IV. MULTILINGUAL AND CODE MIXED NLP TECHNIQUES

The mix of code and multiple languages together with pupils switching between English and regional languages leads to the creation of informal and transliterated expressions that must be handled by educational chatbots in Indian contexts.

MULTILINGUAL NLP TECHNIQUES

A single model that works for multiple languages called multilingual NLP enables comprehension for various languages. The two fundamental strategies for this field are

Approach	Description	Limitation / Strengths
Translation-Based Systems	Converts query to English → processes → translates back	Simple, but loses context/idioms and produces cultural mismatch
Natively Multilingual Models	Models trained simultaneously on many languages with shared embeddings	Better semantic accuracy and cross-lingual transfer

Examples: mBERT, XLM-R, mT5, LaBSE, LLaMA-2 multilingual.

CODE-MIXED NLP

Code-mixed language forms occur in sentences because of language words we combine such as the following:

- SPA Exam fees kab bharne hai?
- Scholarship ke forms open ho chuke hai kya?

The combination of language switches inside sentences together with a spell check leads to these writing issues.
The following are the techniques used to achieve this:

- Language token detection
- Standardization of database of transliterated words
- Implementation of BPE Sub-word embedding method
- Creation of training code-mixed datasets
- Intent models of context dependent on contextual factors

MULTILINGUAL EMBEDDINGS

Meaningful vector forms are formed from text through embedding methods. With shared multi-language embeddings semantic similarity between languages becomes traceable so that
“admission deadline” ≈ “admission ki last date” ≈ “admission mudivu thedi”

RETRIEVAL-AUGMENTED GENERATION (RAG)

RAG uses institutional document references to generate chatbot responses instead of randomly making answers from standard data.
User question → document embedding semantic search → relevant document retrieval → LLM response generation based on documents.
It gives better results while minimizing fake results.

CONTEXT MANAGEMENT
To produce fluid dialogue exchange the chatbot has to continue recording:

- Language choices of users
- The latest conversations
- Items such as dates and course details and fees

Methodological approaches feature sliding conversation windows and entity memory and session summaries.

HUMAN OVERRIDES

Certain requests need manual operation (e.g., complaint handling as well as exception to workplace policy rules). A trustworthy chatbot must:

- Recognize reduced confidence levels
- Mark conversations needing attention
- Connect to human lead agent

V. EXISTING SYSTEMS AND RESEARCH STUDIES LITERATURE REVIEW

Multilingual and code-mixed chatbots received research attention across businesses such as customer support, conversational AI, rural education and multilingual personalization. The number of educational chatbot studies remains low while those applying document-grounded response generation with code-mixed understanding in multiple languages are extremely rare. We analyze the objectives, methods, results and restrictions of 10 literature publications.

[1] Orosoo et al., 2024 — Multilingual Chatbots Semantic Enhancement
Goal: Advance multilingual semantic understanding to enable smooth cross-cultural chatbot communication.
Methods Used: Implemented transformer-based multilingual embeddings to enhance semantic similarity scoring.
Results: Found high accuracy in cross-cultural intent classification with pure language inputs. Limitations:
• No support for mixed-language inputs
• No RAG or document reference
• No retention of prior contexts

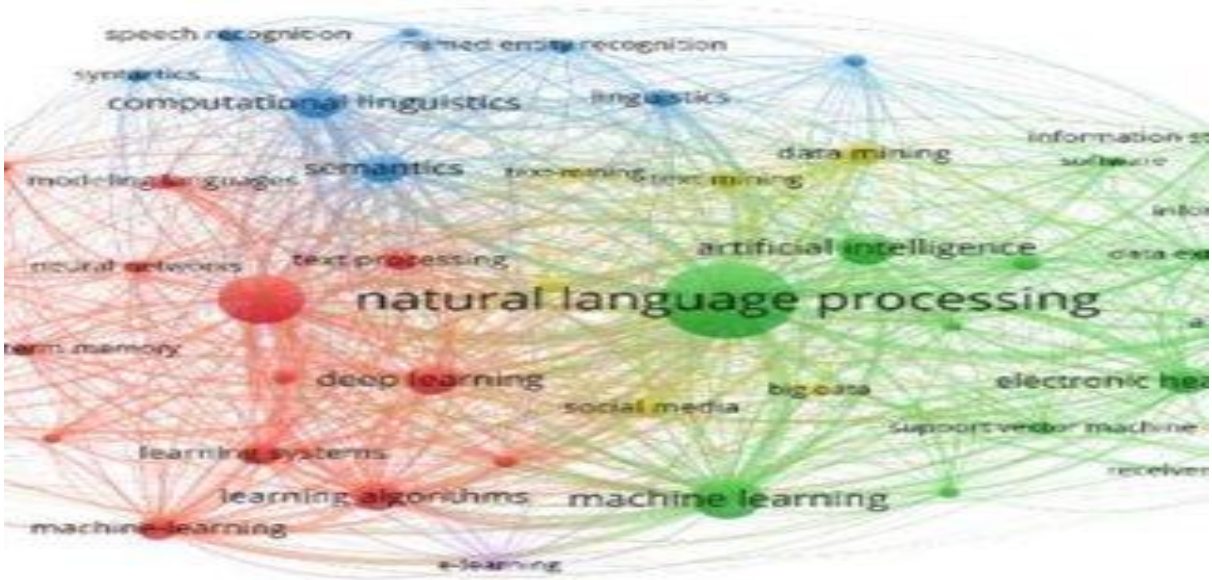

Fig. 3: Multilingual Input → Transformer → Intent Classification → Response

[2] Chethan & Preethi, 2024 — Multilingual Chatbot Educates Rural Users Through AI
Goal: Develop a chatbot that supports rural students by providing assistance through several regional languages.
Approach: Founded on rule-based answers and multilingual NLP processing structured FAQ content. As a result, non-English speakers gained improved access to services.
Limitations:
• The forum is form-bound
• A stock standard answers only
• System does not allow admin document upload and has no dynamic dataset

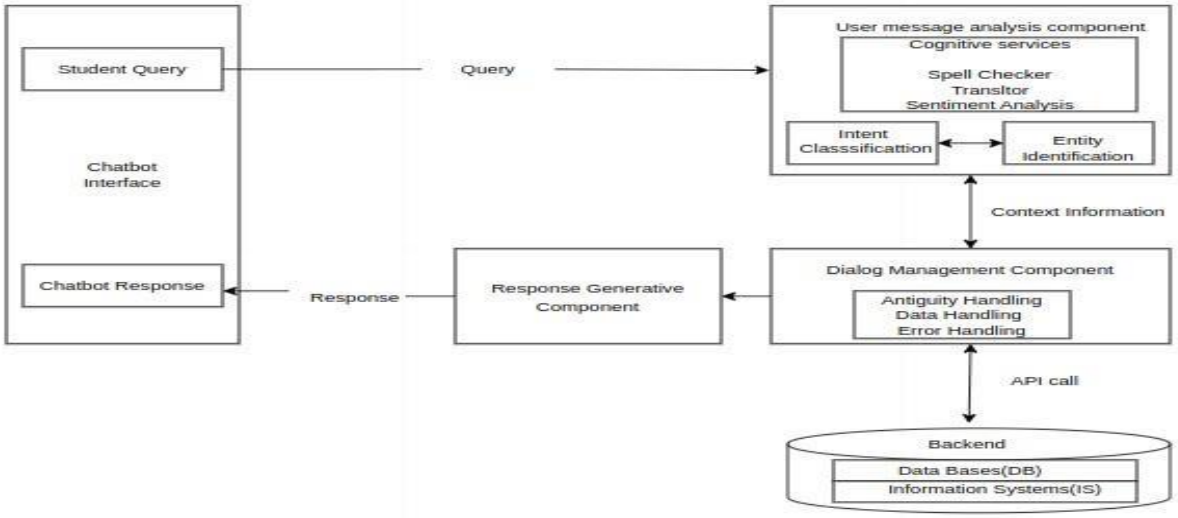

Fig. 4: User - FAQ Lookup - Template Response

[3] Singh et al., 2023 — Multilingual Indian Languages Chatbot
Aim: Provide access to chatbot features in every Indian regional language through use of translation API.
Methodology: Hindi → English → Model → English → Hindi character translation chain.
Result: Constructed a working system with heavy dependence on translation accuracy.
Limitations:
• Vanishing meaning and cultural essence through translation
• Incorrect response generation as a result of broken translation chain
• Hinglish and other hybrid Indian languages remain unsupported

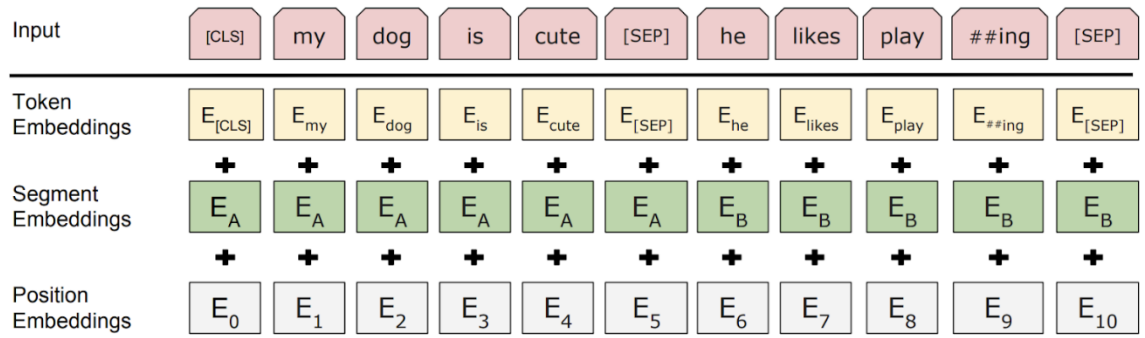

Fig. 5: Hindi → English → Model → English → Hindi

[4] Meta AI, 2023 — LLaMA-2 Multilingual Foundation Model
Goal: Build intellectually capable LLMs that show multi-language reasoning abilities.
Methodology: Large multilingual datasets pre-training; text generation leading-edge technology.
Result: Convincing multi-lingual performance with excellent contextual understanding.
Limitations:
• Need GPU clusters available
• Too heavy to bring along to college
• Without RAG still prone to invent facts

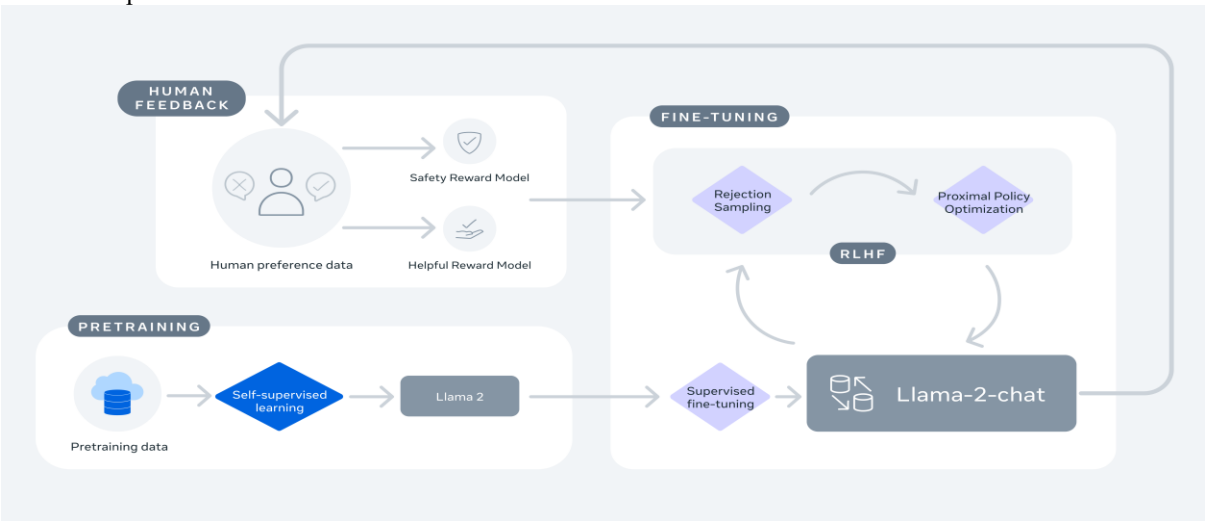

Fig. 6: Prompts - LLM - Fluent Response Without grounding

[5] Rasa Technologies -- 2022 Open-Source Multilingual Conversational AI
Goal: Build modular framework to support business chatbot development.
Methodology: Language-agnostic entity extraction and multi-language intent recognition.
Result: A high degree of developer flexibility through modular pipelines.
Drawbacks:
• Mixed sentences pose challenges for the system
• English and regional languages are handled independently
• Extensive amount of labelled data needed

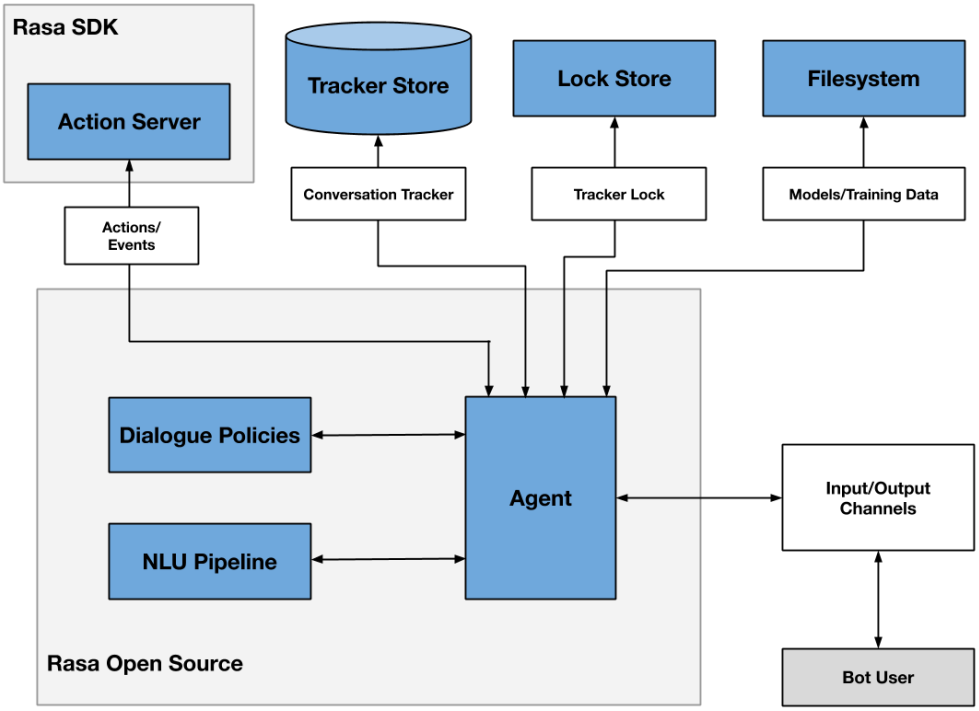

Fig. 7: shows User Intent which moves to Intent Classifier and then produces Pre-programmed Response

[6] Lin et al., 2021 published XPersona: Evaluating Multilingual Personalized Chatbots
Goals: Study a multilingual chatbot that continuously maintains specified user persona and tone through language translation.
Study of a multi-language persona-based data collection.
The results present proof that LLMs maintain personality when converting across different languages.
System Constraints:
• Designed exclusively for non-information systems
• Documents cannot be searched using the current system
• Limited to non-academic applications

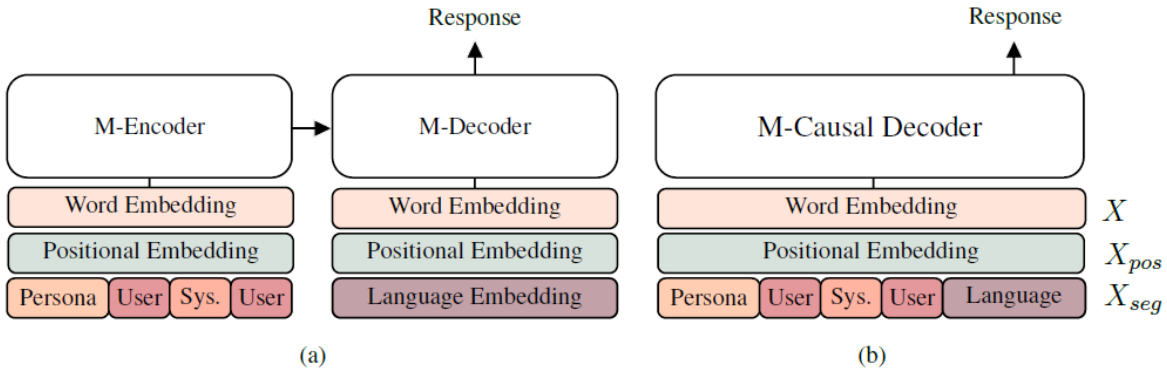

Fig. 8: presents User Preferences flowing to Persona Memory before generating Persona-Aligned Response.

VI. MULITLINGUAL AND CODE MIXED CHATBOTS OBSTACLES

- Presence of code-mixing makes standard grammatical rules obsolete
- Bytes different from each other yet bytes converted from each other
- Available datasets with labels missing in local languages
- Distinctive dialect types and vernacular vocabulary
- Some LLM-generated misinformation risks without RAG mechanism
- Difficulty to maintain conversation flow together with context retention.
- Not able to identify low-confidence answers because detection)

VII. ADVANTAGES AND LIMITATIONS OF EXISTING RESEARCH

ADVANTAGES

- User convenience and accessibility have improved
- Institutional workload has dropped
- Continuous 24-7 accessibility to service
- Speed of response surpasses manual customer support capabilities

LIMITATIONS

- Designed essentially to support English-speaking users
- The system produces weak output for blended-code inputs
- Dependence on institutional document repositories
- Human participation in escalations is not offered by the system
- The system has restricted its Memory functionality for context tracking
- The system has yet to optimize its usage for higher education contexts

VIII. DISCUSSION

Research reveals the transformation of multilingual chatbot architectures from rule-based methods to transformer-based large language models (LLMs). Remote students need more than fluent conversations to receive effective assistance from student support. Institutions have to authenticate accuracy; growing inclusivity demands support for both regional native expressions and code-mixed language; human support escalation remains necessary to ensure system reliability. The forthcoming educational chatbot models will implement hybrid frameworks that unite:

- Multilingual LLM comprehension capabilities
- Fundamental document retrieval capabilities
- Context preservation methodologies
- Oversight by humans

IX. CONCLUSION

This review demonstrates how modern AI chatbot advancements remain unable to provide education communications with a student-centric multilingual chatbot which supports code-mixed content. Dry current systems fall apart because they use fixed intent frameworks, depend on translations only, lack document grounding capabilities, lack context memory and lack human fallback options. The study establishes the importance of creating an educational chatbot which blends multilingual command processing and code-mixed understanding with document verification for delivering combination inclusive academic support to students.

X. ACKNOWLEDGMENTS

We would like to thank our mentor Ms. Harika for helping us with the research process, collecting information, and constructing this paper.

REFERENCES

[1] M. Orosoo, *et al*., “Enhancing NLP in multilingual chatbots,” **2024**.
[2] C. K. Chethan and P. Preethi, “AI-based multilingual chatbot for rural education,” **2024**.
[3] U. Singh, *et al*., “A multilingual chatbot for Indian languages,” **2023**.
[4] Meta AI, “LLaMA-2 multilingual foundation model,” **2023**.
[5] Rasa Technologies, “Rasa open source conversational AI,” **2022**.
[6] Z. Lin, *et al*., “XPersona: Evaluating multilingual personalized chatbots,” **2021**.