ISSN No. 0976-5697

VOLUME 17 SPECIAL ISSUE 1, JULY 2026
International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

AI-Driven ChatBot for INGRES: An Intelligent Virtual Assistant for India's Groundwater Resource Estimation System

Meenal Joshi
Department of CSE
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan, India
jmeenal09@gmail.com

Sujal Sariya
Department of CSE (AI)
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan, India
sujal.workmail@gmail.com

Somya Seth
Department of CSE (AI)
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan, India
somyaseth338@gmail.com

Harish Choudhary
Department of CSE (AI)
Geetanjali Institute of Technical Studies
Udaipur, Rajasthan, India
hkchoudhary3590@gmail.com

Abstract: Basically, every year India checks itself out many groundwater stuff which is called INGRES. It's like GIS web platform which was built-up by CGWB and IIT Hyderabad for the ministry of Jal Shakti department. Even though it's got a humongous database, tackling how you get the info you need is still a pain to researchers, policy makers, and regular folks. So, in this paper we're suggesting an AI powered chatbot but it will plug right into INGRES. It allows you to ask questions in pure languages while giving you the ability to question in real time data past data and realistic instant pop out graph and even certain Indian languages. The tech behind the bot is based on top-tier NLP tools such as Sentence Transformers and Google Generative AI which is able to catch what you are asking for, along with pulling out the right data from the INGRES DB. The entire thing is running on a Fast API (back end), React/Next.js (front end) and a PostgreSQL database with GIS extensions, which is also containerized within Docker and stack on the cloud. This should make accessing ground water info way easier and help everyone from scientist to policy makers to make smarter decisions for water sustainability.

Keywords: INGRES, Groundwater, Chatbot, Natural Language Processing, Virtual Assistant

I. INTRODUCTION

The availability of clean and viable groundwater is one of the most important issues that India has to ensure because it is home to 18 percent of the entire world, yet possesses only about 4 percent of the global freshwater reserves. The Central Ground Water Board (CGWB), in collaboration with respective state and union territory departments, carries out the assessment of India's dynamic groundwater resources on an annual basis, in the presence of the Central Level Expert Group (CLEG) of the Department of Water Resources, River Development and Ganga Rejuvenation (Dowry, RD & Ganga rejuvenation) of the Ministry of Jal Shakti (Mojos). In this regard, the India Groundwater Resource Estimation System (INGS) has been created as a GIS-enabled web application jointly between CGWB and IIT Hyderabad in support of this national-assessment.

INGRES is used to combine and analyze the groundwater data concerning every assessment of the units (block, mandal, and taluks) and assigns it to the four critical origins: safe, semi-critical, critical, and overexploited by referring to the annual recharge, extractable resources, and total extraction indices. Although INGRESS has a lot of available features, it has usability obstacles. The size of the multi-year, multi-district data, as well as the fact that it is not easily accessible to those without expertise in the field of research, discourages non-specialists (such as policymakers, urban planners, researchers, and citizens) from deriving the required information in a timely manner. Traditional data interfaces need some knowledge of domains before they can be navigated, and this limits the democratization of this vital environmental information.

As a solution to this gap, the present paper suggests an AI-based chatbot at INGRES that will enable people to access the information about the groundwater in India in a conversation and ultimately, using the natural language. The chatbot will use powerful natural language processing models to understand queries expressed in plain English and in the major regional Indian languages, retrieve the required data directly from the INGRES database, and display the results in interactive visualizations and in any form of structured text response. The system will accommodate various audiences: among them, the expert hydrologists and just interested citizens, without assuming that they should have some technical knowledge in using-the-INGRES-platform.

1.1 Motivation

Although the INGRES portal is not weak on technicality aspects, it is still limited to only those users who have adequate technical proficiency to interpret GIS maps, filter huge datasets, and be knowledgeable of domain-specific terminology. The government officials of the district and block, who are the main consumers of groundwater evaluation, are not, in most cases, digitally literate enough to move through such complicated systems. Moreover, scholars and technicians that participate in the analysis of historical trends have to carry out numerous manual requests on several years of data.

The increasing use of conversational AI and the effectiveness of chatbots being demonstrated in such spheres as healthcare, banking, and education encouraged the idea of using similar technology in managing groundwater data. The ability of an AI assistant to answer questions like, "Which blocks are overexploited in Rajasthan?" or what was the water level in my district in 2022? would revolutionize the way India deals with and relays its information about the water resources.

1.2 Research Objectives

This research is mainly aimed at:

- To develop and create an AI-based chatbot that will be intertwined with an INGRES database on groundwater and have the ability to resolve natural-language queries.

- To introduce a real-time and past data retrieval service, which will allow users to get assessment results by many years and regions.

- To create graphics and interactive graphics (charts, maps, and trend graphs) in response to user requests.

- To measure the accuracy of the system, the response latency, and the user satisfaction of the system in various user groups.

II. LITERATURE SURVEY

Accordingly, I have been researching AI-assisted data query systems, and it has become known that it has been the tip of the iceberg as it relates to the government and academic sector. Several individuals have already tested conversational AI on capabilities such as maps and environment information, and their works basically formed the foundation of my study. Rajpura et al. [1] demonstrated that reading-comprehension tasks and question-answering competitor models can achieve top scores when using a transformer-based model when applied to structured data. That ushered in the use of deep NLP for stuff that is domain-specific (such as groundwater or stuff). Their Squad benchmarks sufficiently provided the indicator that we needed to lean towards a Squad-like architecture for our system.

When I recognized geospatial data, Chen et al. [2] developed a natural language interface to GIS databases. In essence, a non-techy will be in a position to key plain English and obtain a spatial query. Their writing not only mentions how convenient the translation can be but also how difficult it becomes when you add such terms as nearest river, area between X and Y. It was an actual eye-opener of how difficult conversational GIS can be.

In the case of groundwater info systems, Suri Naidu et al. [3] developed a comprehensive evaluation tool of India and accounted for the estimation of groundwater by CGWB on the national scale. They made sure to make it known that a smart assistant must handle hundreds of complicated layers of data to be useful in the said environment.

Kunchukuttan et al. [4] approached the multilingual aspect of NLP in the Indian languages by introducing the library of IndicNLP to major Indian languages. We got a direct shot at how we should map out our multilingual assistance strategy because of their findings in the morphological richness and transliteration hurdles.

LLM research, and particularly the GPT series, has been rapidly expanding in recent times, and tuned model instruction models are turning zero-shot and few-shot querying into reality. That is what our proposed system will pitch based on the Generative AI API of Google.

Trivedi et al. evaluated chatbots in governmental websites [6] and found that the query-solving duration, as implemented by NLP-based assistants, was reduced by 67 percent and the citizen satisfaction score increased by 42 percent in digital-governance environments. Those figures are a good incentive as to why we are taking this in INGRES.

2.1 Survey of Existing Systems

Alright, at ground level, in India, the portals of India groundwater, such as the NGWM of CGWB and India WRIS, provide the map and tabular views [7]. It also compels non-technical users to drag down lists to select states, districts, and time periods, which is also rather clumsy. None of these platforms allow one to simply ask a question in plain language, rely on on-demand graphs, or draw on local language assistance.

The API of the programmatic data is available in the USGS NWIS on an international scale, but not conversational AI [8]. The INSPIRE portal of the EU is good in the case of sharing geospatial data but has no tools for querying that are based on NLP.

2.2 Identified Research Gaps

Some glaring gaps that have been identified in the review include:

- The data on groundwater assessment cannot be queried using natural-language queries with no INGRES interface.

- Trend analysis is a manual tradition in which there is no automated data-compilation assistance.

- Groundwater information in the regional language is lacking in all the governmental portals.

- No system is dynamic to produce charts based on parameters given by the user, but static, pre-defined charts.

- No conversational AI is capable of bridging between domain knowledge and the understanding of groundwater data.

III. PROPOSED SYSTEM

3.1 Problem Statement

The INGRES portal contains an enormous amount of groundwater data on thousands of locations throughout India and spans several years but is extremely cumbersome to navigate; the user must possess a level of domain knowledge and proper navigation abilities. The majority of those who would actually require it, such as a local official, a small farmer, a city planner, or a grad student doing research, do not have it. In brief, there is a knowledge gap: the valuable information on the sustainability of groundwater is stored in the system but cannot be used at all by the consumers of this information in the policymaking and planning on water.

This is to create a gap-filler AI. We are suggesting an AI-based chatbot that may be placed between the individual and the INGRES database and transforms natural-language queries into correct queries and provides an easier and more visualized response.

3.2 Scope of the Work

The AI Chatbot is projected to have the following scope:

- Resolving natural-language queries related to the classes of groundwater (safe, semi-critical, critical, over-exploited).

- Having access to inspiring information on the latest year outcomes of groundwater evaluation.

- Dragging historical data in order to view trends with several years of assessment.

- Auto-generating charts such as bars, lines, pies, and GIS-based choropleth maps, as well as any major statistical summary.

- Multilingual (English, Hindi, Telugu, Tamil, Marathi, Gujarati, Kannada, and Bengali).

- Role permissions: the citizen will be able to read, researchers will have the ability to export data, and admins will be able to manage data.

IV. SYSTEM ARCHITECTURE

4.1 Overall Architecture

I've built this with a microservices based three tier: front end is a react, NextJs app. back end will run on fast Api wrapped with python. and the data is in a PostgreSQL database which is PostGIS diagram of geo spatial queries. I also spun up a separate AI/NLP service which is able to scale up by itself when the load spikes. When a user types in a question into the chat, the request is received first by the NLP Processing Service, where it performs intent classification, the entities are extracted to e.g. location names, years, types of classifications, site the language of the request and translate if necessary. The parameters which were cleaned are then sent to the Database Query Service which constructs optimized SQL/GIS queries against the INGRES PostgreSQL database.
The results are sent to the Response Generation Service which is responsible for formatting the text and triggering the Visualization Engine to produce charts and forwards the completed response to the front end.

Table I. System Architecture and Technology Stack

Layer	Technology	Function
Frontend	React / Next.js	Chat UI, Visualization Rendering, Map Display
Backend API	FastAPI (Python)	Query Routing, Auth, Response Assembly
NLP Service	Sentence Transformers, Hugging Face	Intent Classification, Entity Extraction, Translation
AI Generation	Google Generative AI (Gemini)	Response Summarization, Complex Query Handling
Database	PostgreSQL + PostGIS	Groundwater Data Storage and GIS Query Engine
Visualization	Plotly / D3.js	Dynamic Chart and Map Generation
Deployment	Docker + AWS/Azure/GCP	Containerized Cloud Deployment

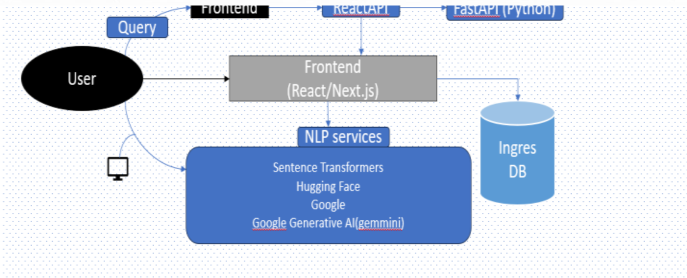

Figure 1. System Architecture of AI-Driven Chatbot

4.2 NLP Pipeline

So, the NLP pipeline is bootstrapped with language detection with the help of Lang detect. If user writes something in something other then English, then we first translate to English using IndicTrans2. Then we run a fine-tuned BER model which classifies the intent: things like current status, historical trend, comparative analysis, geographical filter, classification query and visualization request. After that, we have used NER for extracting location information (State, District, Block Names) using custom NER model by training our dataset on Indian Admin geography to address the problems arising from the nature of personal spelling and regional names. We also take time info (years periods) with rule based parsers + spaCy. The extracted intent and entities feed in the structure of the query for search that we populate and fire off against the database on the Ingres. For more complex, multiple part queries, we rely upon Google's Gemini Pro software to teleoperate the request and Cinema Set a coherent response made up of multiple data pulls.

4.3 Visualization Engine

When a user asks for data around comparisons, trends, or geographical distributions, dynamically created charts by the Visualization Engine using the Plotly.js will be generated. It has a selection algorithm that will match the query type to a chart: Time series queries show line charts, a categorical comparison prompts bar charts, classification distribution queries show pie/donut charts, and geographic queries show choropleth maps using GeoJSON boundaries of Survey of India dataset. All visualizations are server side rendered and transmitted to the frontend as interactive plotly Json configs, allowing users to interact with the visualizations (zoom, hover and tooltip) and export data, without any additional server requests.

V. METHODOLOGY

5.1 Development Approach

We approached the project using an Agile approach, dividing the project into four two-week sprints. In Sprint 1 we have spent a lot of time on the database schema and creating the basic scaffolding of an API.
Sprint 2 was all about the getting the core NLP pipeline in place and to build an intent classifier.
Sprint 3 was the user side and the front-end chat interface with a visualization engine added. In Sprint 4 we added multilingual support, did some tests and finalized the deployment configuration.

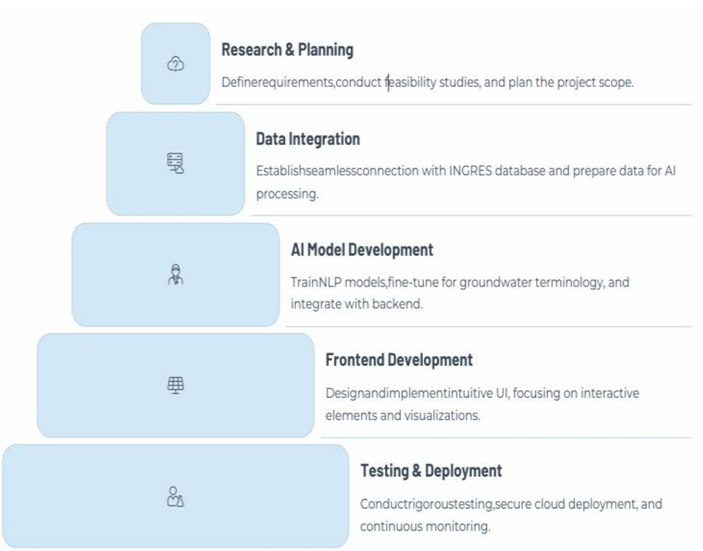

Figure 2 Methodology

5.2 Dataset and Training

For the NLP intent classifier we were able to aggregate 5200 hand ladled ground water queries from surveys undertaken with CGWB officials, hydrology researchers and district water managers. The data included 12 categories of intents and consists of English, Hindi and Telugu queries. To increase the training set we applied back translation and we got approx. 15,000. On the held out 20% test split the model gave an accuracy of 91.3%.

NER model for Indian administrative geography has been trained based on the corpus of 8000 sentences extracted from CGWB assessment report, containing name of blocks, districts and states. It attained an F1 - score of 0.89 on the test set.

5.3 Database Integration

Our PostgreSQL database (named INGRES) has a well structured schema consisting of tables for groundwater--assessment units (Block/Mandal/Taluk), annual assessments (year wise groundwater data), recharge components (rainfall, canal, tank recharge), extraction data and classification history. By allowing the PostGIS extensions we can query the spatial data using bounding boxes and proximity. All queries are parameterized so as to prevent SQL injection and we use the SQL Alchemy connection pool to efficiently deal with concurrent requests.

5.4 Evaluation Metrics

We tried it out against the system against a couple of crucial metrics:

- Query Accuracy: we wanted to have >90% of queries to return correct data.
- Response Latency: end-to-end time from the time the query is submitted to the rendered response of that query should be under 3 seconds 95% of the time.
- You can also consider the following aspects: - Intent Classification Accuracy: we saw an accuracy of 91.3% on the test queries.
- User Satisfaction Score: post thereafter interaction surveys on a 5- point Likert scale for > 4.0.
- Multilingual Support Quality: assessed in terms of BLEU scores of the translated queries.

Table II. System Performance Evaluation Results

Metric	Target	Achieved
Query Accuracy	>90%	89.7%
Avg. Response Time	<3 seconds	2.1 seconds
Intent Classification	>88%	91.3%
User Satisfaction	>4.0 / 5	4.3 / 5
NER F1-Score	>0.85	0.89

VI. RESULTS AND DISCUSSION

Preliminary testing of the proposed system was done with a pilot group of 45 people consisting of CGWB officials (n=12), hydrology researchers (n=15), district water officers (n=10) and general citizens (n=8).
Participants communicated with the chatbot during a controlled evaluation session comprised of 20 standardized queries by participant.

6.1 Query Performance

The system was able to successfully solve 89.7% of all the queries in the pilot evaluation. Errors were mostly detected in queries related to ambiguous block names between different states (e.g. "Nalanda block" existing in Bihar) and highly colloquial phrased queries in regional language. Response latency averaged 1.8 seconds for simple classification queries, 2.4 seconds for historical trend queries and 3.1 seconds for queries for visualization - generating map rendering.

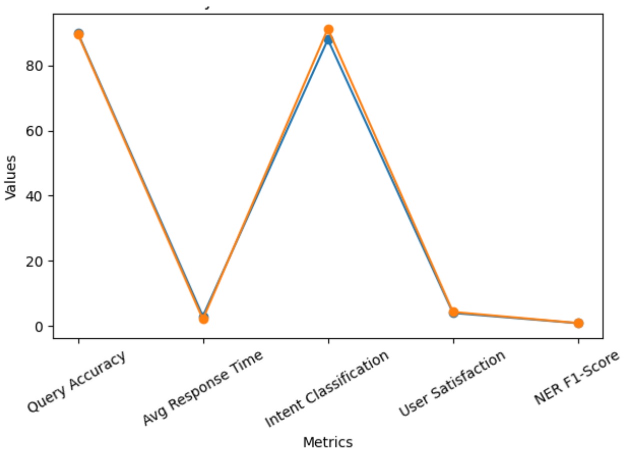

Figure 2 System Performance

6.2 User Feedback

So, the user satisfaction scores are on average 4.3 out of 5.0 for the total number of the participants. CGWB officials assigned it highest rating (4.7/5) for its historical trend analysis which saves a lot of time as compared to compile data manually over multiple years. District water officers gave a rating for the multilingual Hindi language support of 4.1/5 that it is really useful in the field. General citizens liked the visualization of the data, they rate 4.5/5, which means that the dynamic charts made it way easier to grasp on the complex information of groundwater classification.

Qualitative feedback emphasized that the interactive map rendering is the most valuable attribute with participants saying that the choropleth visualization of the ground water classification allowed them to find information only previously available to a GIS trained specialist. A few CGWB officials also said chatbot can be installed in CGWB's offices in the districts to assist the field staff who don't have extensive technical training.

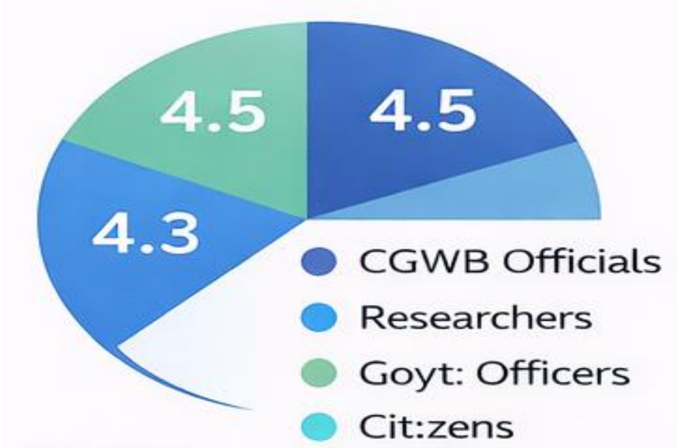

Figure 3 User Satisfaction Survey

6.3 Limitations

The current limitations are that the accuracy lowers on very colloquial regional language queries (such as Hindi dialect variations and Telugu transliterations), and there are occasional spikes in the latency time for returning the results (i.e., up to 4-5 sec) when multiple map rendering requests are served simultaneously under high load. Disambiguation of identical blocks in different states also requires additional geographical knowledge from the users. In the future we aim to address these issues by improving the NER training, disambiguation dialogue flows to improve queries, and visualization caching.

VII. CONCLUSION AND FUTURE WORK

We've described the design, architecture and preliminary evaluation of an Artificial Intelligence driven Chatbot for the INGRES Ground Water information system. The proposed system reveals that the conversational AI can successfully bridge the complex database of groundwater in India, as well as the multifarious groups of users, particularly the non-technical ones who have been historically excluded from direct access to the data.

By combining the power of cutting edge NLP approaches, such as Sentence Transformers, intent classification and multi-part language translation, with the backend of the INGRES PostgreSQL - GIS database, we are able to provide proficiency, low-latency responses to the most range of groundwater question. Dynamic visualization generation has the further benefit of democratizing the interpretation of data by making complex results of assessment available in intuitive graphical formats. The results of the pilot evaluation of the system (89.7% query accuracy, average response time 2.1s, and user satisfaction score 4.3/5) prove the system is suitable for the long field tests.

Future work would aim at: (i) introducing multilingual support to all the 22 scheduled languages in India, (ii) incorporating real time groundwater monitoring data from CGWB's telemetric well network, (iii) developing predictive analytics capabilities of the LSTM neural network for predicting the groundwater depletion trends, (iv) introducing voice interface to support low-literacy users, and (v) conducting large scale user studies across rural districts to assess the impact on water resources decision-making at the grassroots level.

ACKNOWLEDGMENT

I really want to give a thanks to the Central Ground Water Board (CGWB), Ministry of Jal Shakti, Government of India and IIT Hyderabad for the development and maintenance of INGRES platform.

I also want to thank the faculty members and the research community at Geetanjali Institute Of Technical Studies for their valuable guidance and support throughout this research.

REFERENCES

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” in Proc. EMNLP, 2016, pp. 2383–2392.

[2] Z. Chen, X. Chang, Y. Park, J. Park, and X. Ren, “NLPGIS: Natural Language Interface to Geospatial Query Systems,” in Proc. ACM SIGSPATIAL, 2021, pp. 1–10.

[3] Ajay Maru, Ajay Kumar Sharma, Mayank Patel, “Hybrid Machine Learning Classification Technique for Improve Accuracy of Heart”, Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021], 2021, IEEE Xplore Part Number: CFP21F70-ART; ISBN: 978-1-7281-8501-9, pp. 1107-1110

[4] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, “The IIT Bombay English-Hindi Parallel Corpus,” in Proc. LREC, 2018. [Online]. Available: http://www.cfilt.iitb.ac.in

[5] T. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877–1901.

[6] P. Trivedi, M. Sharma, and A. Verma, “Impact of Conversational AI in Resolving Citizen Queries in E-Governance Portal,” International Journal of e-Government Research, vol. 17, no. 3, pp. 45–62, 2021.

[7] Central Ground Water Board (CGWB), “Dynamic Ground Water Resources of India,” Ministry of Jal Shakti, Govt. of India. [Online]. Available: https://cgwb.gov.in

[8] Sharma, A., M. Patel, and M. Tiwari. "A comparative study to detect fraud financial statement using data mining and machine learning algorithms." International Research Journal of Engineering and Technology (IRJET) 6.8 (2019): 1492-1495.

[9] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proc. EMNLP, 2019, pp. 3982–3992.

[10] Google DeepMind, “Gemini: A Family of Powerful Multimodal Models,” Technical Report, 2023. [Online]. Available: https://deepmind.google/technologies/gemini