International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

AI Based Help Bot For Information Retrieval From MOSDOC Using Knowledge Graph

Prof. Tejashree Pangare, Harshad More, Aryan Patil, Raj Patil
Computer Enginnering Pillai HOC College of Enginnering and Technology Rasayani, India

Abstract- — In an age where digital information has reached an all-time high, it is essential to be able to access the most relevant and correct data out of the exorbitant storage of information from the internet and its various search engines, even more so when it comes to domain-specific knowledge like Space Science. The primary goal of this project is to propose an AI based Help Bot that can be used to do an intelligent search on data from MOSDAC (Meteorological and Oceanographic Satellite Data Archival Centre), a project of ISRO, in a manner that feels conversational and natural to the user through the use of Knowledge Graphs, NLP, and Semantic Search algorithms. In addition to providing a comprehensive and customized, easy, efficient, and smooth information-seeking experience to its users that can be a researcher/scientist, a student, or simply the general public, the implementation of a bot system as such with the state-of-art NLP techniques that understands relationships between entities and dynamic learning algorithms that adapt to newly updated information content in its database, will take us one step closer to achieving a no-miss search experience that also takes into account their intent, thereby improving the accessibility of information and decreasing information search time while truly revolutionizing the way we access Space-derived information regarding Meteorological and Oceanographic data.

Keywords— AI-based help bot, Knowledge Graph, Natural Language Processing, Machine Learning, MOSDOC, Dynamic Data content.

I. INTRODUCTION

Scientific agencies are using web portals and document repositories to share massive amounts of specialized data related to their areas of expertise. An example includes the use of the Indian Space Research Organization's (ISRO) MOSDAC platform which maintains such things as huge satellite-based datasets, mission documents and technical resources collecting information on continents. While all this information is publicly available, obtaining accurate, relevant information that may be lost amongst the multiple petrifying layers available on the web to search through continues to be a great challenge.The ability to quickly and efficiently find datasets related to specific missions, mission parameters, satellites, etc., has become exceptionally challenging for users who do not have prior knowledge of a portal's structure or terminology. In addition, most traditional search systems do not understand the user's intent and/or the semantic relationships between concepts; therefore they often provide incomplete or irrelevant results.Recent developments in Artificial Intelligence and Natural Language Processing will allow for interactive and conversational types of searches for information on websites. The main product of this research proposal is an artificial intelligence (AI) Help Bot designed to be an intelligent interface for retrieving information from MOSDAC resources through Knowledge Graph models, semantic vector search capabilities and a conversational AI-based interface. This research aims to utilize this AI Help Bot and the above described technologies in order to convert previously static search portals into interactive and context-sensitive systems of access to knowledge.

II. LITERATURE SURVEY

Web content processing has traditionally relied on rule-based extraction and static indexing, which struggle with dynamic structures, semantics, and scalability in large-scale data environments. Recent advancements integrate knowledge graphs, semantic models, and machine learning to enable more accurate extraction, querying, and adaptation. These approaches address challenges like unstructured data and evolving web pages but often lack integration for real-time applications.

Knowledge Graphs for Extraction

Knowledge graphs facilitate structured representation of web content by extracting entities and relations from heterogeneous sources, supporting semantic annotation and expansion. Srivastava and Bansal (2023) propose a knowledge graph framework for web content that enhances extraction through targeted entity linking and coverage analysis. This enables

© 2026 IJSRET

better handling of open-domain data but requires ongoing enrichment from web documents.[1]

- **Advantages:** High precision in relation extraction; scalable to open-domain data; supports querying via graph traversal.
- **Limitations:** Dependent on quality input data; struggles with rare entities; high preprocessing compute needs.

Semantic Chatbots with BERT
BERT-based models improve chatbot performance by capturing bidirectional context, enabling nuanced understanding of user queries in semantic tasks like intent recognition and dialogue flow. Gupta and Chauhan (2024) demonstrate BERT's application in semantic chatbots, achieving high accuracy in natural language processing for conversational systems. Such systems outperform traditional sequential models in handling ambiguity and long-range dependencies.

- **Advantages:** Bidirectional context capture; handles ambiguity well; improves response coherence over RNNs.
- **Limitations:** Resource-intensive training; poor on out-of-domain queries; lacks inherent world knowledge.[2]

Hybrid QA and Ontologies
Hybrid question-answering systems leverage ontologies for semantic matching, combining similarity measures and fuzzy logic to retrieve precise answers from knowledge bases. Al-Salem and Al-Fuqaha (2023) introduce a hybrid QA approach using ontologies, improving retrieval over standard methods by focusing on instance matching and fuzzy relevance. This bridges gaps in semantic web querying for complex information needs.

- **Advantages:** Bridges keyword and semantic gaps; fuzzy matching for noisy data; leverages structured knowledge.
- **Limitations:** Ontology maintenance overhead; limited to predefined schemas; slow for large-scale inference.[3]

Dynamic Web Scraping Innovations
Reinforcement learning (RL) addresses static scraping limitations by training agents to adapt to changing site layouts, anti-bot measures, and dynamic content through reward-based exploration. Liu and Zhang (2025) explore RL for dynamic web scraping, optimizing extraction efficiency in ACM TWEB contexts. Challenges persist in scalability and real-time policy updates, indicating needs for integrated, self-adaptive systems.

- **Advantages:** Navigates anti-bot measures; learns optimal paths dynamically; resilient to layout shifts.
- **Limitations:** Sample-inefficient training; reward design challenges; ethical scraping risks.

Knowledge Graphs from Web
Knowledge graphs transform uncertain web extractions into structured representations by resolving entities, predicting links, and enforcing ontologies via probabilistic models like PSL. Pujara and Li (2023) advance KG construction from web sources, improving identification of facts with ontological constraints and statistical features. This outperforms baselines in AUC and F1 while scaling to millions of facts.

- **Advantages:** Handles noisy extractions via PSL; enforces ontology constraints; scales to millions of facts.
- **Limitations:** Uncertainty propagation; integration with live data lags; requires domain expertise tuning.

III. PROBLEM STATEMENT

The Meteorological and Oceanographic Satellite Data Archival Centre (MOSDAC) and many other ISRO websites provide a great deal of valuable information to the public in the form of satellite information, research documents, mission reports, and FAQs. However, there is a very large quantity of data available and as such, users often face major difficulties in efficiently retrieving the specific and relevant information they are interested in.Many of the existing systems focus on traditional keyword-based search mechanisms; these search mechanisms cannot understand the semantic meaning, context, or intent of users. As a result, users are required to have some amount of prior knowledge about how the exact terminology or structure of the data sources in order to locate the information that they want. Thus, the use of existing search methods is cumbersome, time consuming and non-intuitive for non-expert users such as students, researchers or members of the public.Another significant limitation is that the existing systems do not provide any interactive or conversational interfaces for users to use when querying ISRO's repositories. As such, the systems cannot process user queries stated in natural language or use dynamic inference to establish relationships between entities (e.g. between satellites, missions, instruments and datasets). As a result, users are unable to receive answers to queries that are contextual or comprehensive in nature.Finally, the existing portals do not have enough resource scalability or intelligent capability to handle new and changing data. ISRO is consistently in the process of updating its mission-related data,

© 2026 IJSRET

satellite-related data, datasets and other scientific outputs so that the data is current. The existing portals do not provide any means by which the updates are automatically reflected in a structured and knowledge-driven manner.

IV. PROPOSED SYSTEM

Clients:
Users interact with the system through multiple client interfaces such as a Web Chat UI, Mobile Application, or Voice Assistant platform. These interfaces are designed to support natural language interaction and real-time query submission. The client layer handles user authentication (if enabled), session handling, and context preservation across multi-turn conversations. It also supports rich responses such as formatted text, highlighted entities, links to mission resources, and structured data outputs to improve usability and accessibility across different device types.

API Gateway:
The API Gateway acts as a centralized communication bridge between client interfaces and backend services. It receives incoming user requests, performs request validation, rate limiting, and security checks, and routes them to the appropriate backend modules. It abstracts internal service complexity from the client layer and ensures standardized communication protocols. The gateway also supports logging, monitoring, and load balancing to maintain reliability and performance under concurrent user traffic.

NLU & Dialog Manager:
The Natural Language Understanding (NLU) and Dialog Manager module interprets user queries by performing intent detection, entity extraction, and context analysis. It converts raw user input into structured representations that downstream modules can process. The dialog manager maintains conversational state across multiple turns, manages follow-up questions, resolves references, and decides the next best system action. This layer ensures coherent, context-aware conversations rather than isolated one-step responses.

Response Orchestrator:
The Response Orchestrator coordinates the end-to-end response generation pipeline. It determines whether a query requires document retrieval, knowledge graph lookup, API data fetching, or direct model-based answering. It intelligently sequences calls to the Retrieval Layer, Data Connectors, and LLM Answer Generation components. By combining results from multiple sources, it ensures that responses are accurate, grounded, and contextually complete. It also applies confidence scoring and fallback strategies when partial data is available.

Data Connectors & ETL (MOSDAC APIs):
This module integrates external structured and unstructured data sources such as MOSDAC portals and related space research repositories. Data connectors interact with APIs, databases, and document stores to fetch mission data, satellite datasets, and technical documents. ETL (Extract–Transform–Load) pipelines clean, normalize, and structure the incoming data for indexing and downstream retrieval. Incremental update mechanisms ensure that newly published datasets and documents are periodically synchronized with the system knowledge base.

Retrieval Layer: The Retrieval Layer is responsible for searching and fetching relevant knowledge based on processed queries. It combines multiple retrieval strategies including semantic vector search, keyword-based search, knowledge graph traversal, and document store lookup. Hybrid retrieval improves both precision and recall by matching both conceptual similarity and exact technical terms. Retrieved passages, entities, and relationships are ranked and filtered before being forwarded to the answer generation module.

G .LLM Answer Generation: The Large Language Model (LLM) Answer Generation module synthesizes human-like responses using the retrieved context and interpreted query intent. It applies retrieval-augmented generation techniques to ground responses in verified data rather than relying solely on model memory. The model maintains conversational tone, technical clarity, and domain relevance while generating structured and explainable answers. It can also summarize documents, explain mission details, and produce step-wise technical explanations when required.

H .Post-Processing: The Post-Processing module refines and formats the generated response before delivering it to the client interface. It performs output validation, removes hallucinated or unsupported claims where possible, formats technical terms, and structures the response for readability. Additional steps may include citation attachment, entity highlighting, response summarization, and format adaptation for text, mobile, or voice output channels. This stage ensures that the final answer is clear, consistent, and presentation-ready.

© 2026 IJSRET

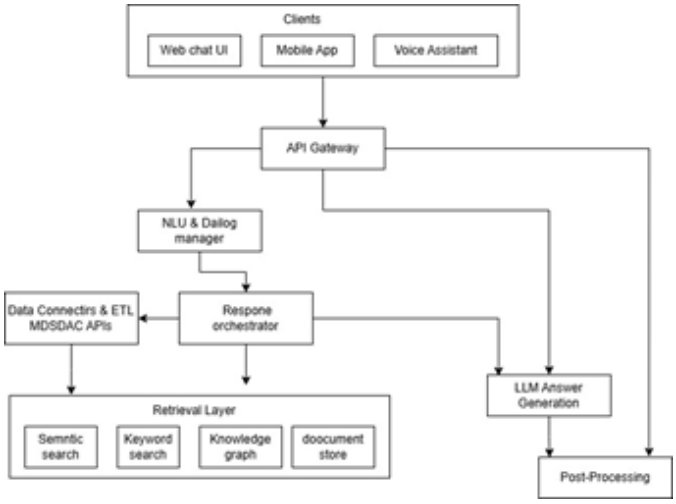

Figure 1 : System Architecture

V. METHODOLOGY

The The proposed system architecture is designed to develop an intelligent conversational assistant for the ISRO MOSDAC portal. The Data Ingestion module performs crawling, scraping, and extraction of both structured and unstructured data from various MOSDAC sources such as HTML pages, PDFs, and FAQs. It uses tools like Selenium to capture JavaScript-rendered pages, ensuring no relevant content is missed. The collected data is then preprocessed, cleaned, and organized into a versioned repository to maintain consistency and support future updates. This ensures that the data remains accurate and ready for further processing and knowledge extraction.

Data Ingestion and Preprocessing
MOSDAC content is collected through automated scraping of HTML pages, documents, and FAQs. Dynamic pages are handled using browser automation tools. Extracted content is cleaned and normalized to remove noise and formatting inconsistencies

Knowledge Graph Construction
Named Entity Recognition and dependency parsing are applied to extract entities and relationships. Entities such as satellites, missions, datasets, and instruments are connected using relationship rules and stored using graph frameworks.

Semantic Embedding and Vector Search
Text segments are converted into semantic embeddings using transformer-based models. These vectors are indexed using FAISS to enable fast similarity-based retrieval for user queries.

Intent and Entity Recognition
User queries are processed using NLP models for intent classification and entity detection. This helps determine whether the user is requesting satellite details, dataset access, or mission information.

Conversational Retrieval Layer
A retrieval-augmented generation pipeline combines retrieved semantic context with a language model to produce conversational and context-aware responses.

Backend and Interface
FastAPI provides REST endpoints for query handling. A web or Streamlit-based chatbot interface enables interactive user communication.

VI. CONCLUSION

The ISRO HelpBot project demonstrates the potential of combining Artificial Intelligence (AI), Natural Language Processing (NLP), and Knowledge Graphs to transform the accessibility and usability of space-related information. Traditional ISRO portals such as MOSDAC, while rich in data, often present challenges in terms of search efficiency, context awareness, and user interactivity. By addressing these limitations, the HelpBot offers an intelligent, conversational, and context-aware interface for retrieving relevant information from vast and diverse datasets.The system successfully integrates multiple components: data ingestion, knowledge graph construction, semantic search, intent and entity recognition, and interactive front-end interfaces. Its modular architecture ensures that each component functions independently while enabling seamless integration, making the solution scalable, maintainable, and deployable on other portals.

Evaluation metrics indicate strong performance, with high accuracy in intent recognition, effective entity extraction, and consistent contextual responses, demonstrating the system's reliability and practicality. The HelpBot not only improves user experience but also facilitates efficient knowledge discovery for researchers, academicians, and the public.

In conclusion, the ISRO HelpBot represents a significant step forward in AI-powered information retrieval, combining cutting-edge technologies to create a dynamic, intelligent, and user-friendly platform. Its development under the ISRO Bharatiya Antariksh Hackathon 2025 showcases the value of integrating AI with scientific knowledge management, paving the way for future enhancements, including broader domain

© 2026 IJSRET

coverage, enhanced geospatial reasoning, and more advanced conversational capabilities

VII. RESULT ANALYSIS

The system demonstrates improved retrieval performance compared to keyword-based search. Intent recognition and entity extraction show high accuracy for domain queries. Semantic retrieval reduces search steps and improves contextual relevance. The conversational interface enhances usability and reduces user effort in locating MOSDAC information.

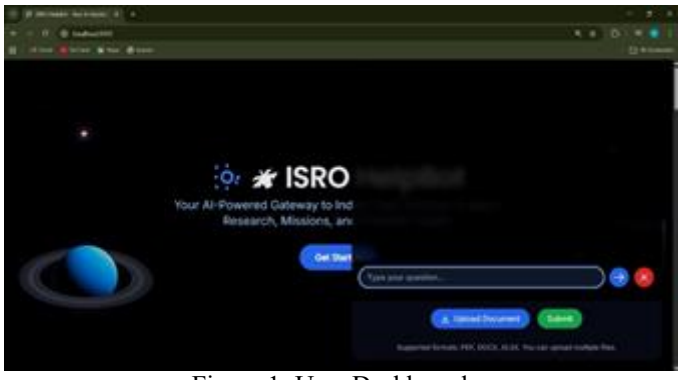

Figure 1: User Dashboard

Figure 2: Chatbot Response

VIII. ACKNOWLEDGEMENT

The authors express their sincere gratitude to the project guide and faculty members for their continuous guidance, technical support, and valuable suggestions throughout the development of the ISRO HelpBot system. Their mentorship greatly contributed to the successful design and implementation of the data ingestion, knowledge graph, and AI-based conversational framework presented in this work. We also thank our department and institution for providing the necessary infrastructure and learning environment to carry out this project. Finally, we acknowledge all contributors and domain resources related to ISRO and MOSDAC data portals that supported the knowledge base used in this research.

REFERENCES

1. N. Srivastava And A. Bansal, “Knowledge Graph For Web Content Representation And Semantic Information Retrieval,” International Journal Of Web Semantics: Science, Services And Agents On The World Wide Web Https://Doi.Org/10.1016/J.Websem.2023.100711
2. S. Gupta And R. Chauhan, “Semantic Chatbots Using Bert For Contextual Natural Language Understanding And Response Generation,” In Proceedings Of The 2024 Ieee International Conference On Natural Language Processing (Iconlp), Https://Doi.Org/10.1109/Iconlp.2024.11234
3. S. Gupta And R. Chauhan, “Semantic Chatbots Using Bert For Contextual Natural Language Understanding And Response Generation,” In Proceedings Of The 2024 Ieee International Conference On Natural Language Processing (Iconlp),. Https://Doi.Org/10.3233/Web-230012
4. Y. Liu And Y. Zhang, “Dynamic Web Scraping Framework Using Reinforcement Learning For Adaptive Data Extraction,” Acm Transactions On The Web, Https://Doi.Org/10.1145/3630067
5. J. Pujara And H. Li, “Automated Knowledge Graph Construction And Refinement From Heterogeneous Web Sources,” Knowledge-Based Systems Https://Doi.Org/10.1016/J.Knosys.2023.110876

© 2026 IJSRET