International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Need of retrieval augmented generation for conversational assistants for government social welfare scheme popularization

Vaidheeswaran Vaishnav [1,*] and Guru Omkar Nathan [2]

[1] *Dalhousie University, Bangalore, Karnataka, India.*
[2] *Gitanjali Olympiad School, Bangalore, Karnataka, India.*

International Journal of Science and Research Archive, 2024, 13(01), 1299–1312

Publication history: Received on 08 August 2024; revised on 16 September 2024; accepted on 19 September 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.1.1747

Abstract

Governments allocate substantial resources to uplift under-privileged communities through various assistance schemes. Despite these efforts, the effectiveness of such initiatives often fails due to variety of reasons including cryptic nature of information available, lack of awareness to the needy people and or lack of transparency. In India, a vast and diverse developing democracy, these challenges are exacerbated by disparities in access to information. To address these issues, we explored the use of conversational AI technologies, including chatbots, which have seen successful application in many developed nations. Conversational AI agents often get challenged to a variety of factual inaccuracies and other types of hallucinations. Our study compares the accuracy of responses across different architectures: Our architectures integrate advancements in natural language processing, such as large language models (LLMs) and retrieval-augmented generation (RAG), along with automatic speech recognition (ASR), knowledge distillation, hybrid retrieval strategies, and text-to-speech systems. Overall this research experiments demonstrate that leveraging retrieval augmented generation using the same model embeddings substantially improves the quality of the output, its similarity to desired responses. This also reduces hallucinations. We also notice that larger models perform better than smaller parameter models.

This paper advocates for continued research in this field, suggesting that further advancements could substantially benefit those the schemes intend to serve.

Keywords: Government Schemes; Conversational AI; Retrieval Augmented Generation; Text-Speech; Hallucination

1. Introduction

Governments worldwide allocate significant resources to uplift marginalized communities, yet the effectiveness of such welfare programs is often undermined by the obscurity and fragmentation of vital information. Government schemes aimed at social protection and poverty reduction have faced significant challenges globally. In VietNam, “Mandatory insurance schemes reach only 32% of the working population and the coverage of voluntary insurance is just 2.1% (1). Similarly, the reliability of government budget execution, a crucial factor for the successful delivery of government programs, shows significant variance globally. Over the period from 2010 to 2021, only slightly more than half of the assessed countries managed to keep their budget execution within 5% of planned expenditures (2).

In poverty reduction, the situation is equally dire may other countries, where ongoing challenges persist in reducing extreme poverty. Based on UN assessments, if the trends continue, about 575 million people will still be living under extreme poverty by 2030 (3). This highlights the slow and uneven progress in poverty reduction, exacerbated by the COVID-19 pandemic which has reversed years of progress (3).

* Corresponding author: Vaidheeswaran, Vaishnav

Despite India's economic growth and a reduction in poverty rates over the years, the implementation of social protection remains inconsistent, often constrained to specific states and lacking a systematic approach (4). Last year, 12 social welfare schemes failed to meet their target in various states (5). In India, a country marked by substantial economic and cultural diversity, this issue is exacerbated by the high rates of illiteracy and digital illiteracy across the needy people. For instance, only 41% of Indian households were covered by any health scheme or insurance by 2019-21 (6)., reflecting a critical gap in accessibility to government support programs. Furthermore, while the internet penetration in rural areas stands at around 37%, the literacy rate is only about 77.7% (6), presenting significant barriers to accessing digital platforms for social welfare.

The failure or slow adoption of government schemes in India is often attributed to several factors that interact together:

- **Lack of Awareness:** There's a significant gap in awareness about government schemes among the target populations. Many people remain unaware of the benefits or how to access them, which leads to underutilization of available resources and programs designed for their welfare. This is often cited by many researchers. (7), (8), (9)
- **Multilingual Diversity**: India's linguistic diversity poses a significant challenge. With numerous languages spoken across different states, government applications primarily in English fail to be comprehensible to many citizens (10). According to the 2001 census, there are approximately 1000 documented languages and dialects in India. (11).
- **Low Literacy Rates**: The low literacy rate further exacerbates the issue, as many individuals in rural areas cannot effectively use e-Governance platforms that are not in their native language (10).
- **Digital Literacy**: Digital literacy refers to the ability to effectively and responsibly use technology to navigate, evaluate, and create or consume information. Apart from general literacy, technology literacy is particularly low, which means even fewer people are capable of navigating e-Governance platforms even if they were available in a more accessible language (10). Despite the push for digitalization under schemes like Digital India, technological barriers still exist, especially in rural and underdeveloped regions where digital literacy is low. This hampers the effective dissemination and utilization of digital-based services (7)
- **Digital Divide**: The digital divide refers to the gap between individuals who have access to modern information and communication technology and those who do not. The digital divide remains a persistent problem, with many rural areas lacking the necessary infrastructure to support e-Governance, such as stable internet connectivity and access to computers or smartphones (10). Poor Infrastructure and lack of steady electricity and the Internet are key causes for e-governance challenges in India (12)

These factors highlight the complex challenges faced in the implementation of government schemes in India, necessitating a multi-pronged approach to tackle these issues comprehensively. While this paper leaves administrative issues out of scope of the research, areas such as “Lack of awareness”, “Technology Barriers”, “Digital Divide” make compelling propositions for overcoming them through recent advancement in technologies.

2. Review of literature

AI chatbots are increasingly collaborating to make information about governmental programs and benefits more popular and accessible. AI-powered chatbots, for instance, assist citizens by answering questions and providing guidance on eligibility and application processes in real-time. For example:

- Dubai Government’s DEWA RAMMA’s service (13)
- Mississippi State Government’s MISSI (14)
- Meet Alex – IP Australia's next-generation virtual assistant (15)
- Alphabot by Singapore Government (16)

2.1. Current State of Chatbot Usage

The deployment of chatbots in public administrations has primarily focused on delivering basic information and advice, which has led to their confinement to simple text-based interactions. This limitation has hindered their potential for creating value within public services, as they often fail to meet the comprehensive service delivery expectations. As a result, government chatbots are generally perceived as tools only suitable for navigating known services and information or handling straightforward inquiries (17)

The adoption of chatbots within the public sector has faced significant challenges, leading to a lack of effective implementation. One primary issue is that many existing chatbot architectures in public administration are not fully

efficient, as they are often built with simple heuristics based conversation managers, designed to respond to pre-defined query patterns. “these chatbots typically rely on basic menu-driven systems that provide information based on pre-selected options” (18). This limited functionality fails to leverage the full potential of AI, resulting in suboptimal user experiences and reduced efficiency. “Additionally, there are concerns about the transparency and comprehensibility of AI technologies, which can hinder trust and confidence in their use” (18). “To address these issues, it is crucial to develop and implement more sophisticated AI-based chatbots that can better simulate human interactions and provide more effective public services”. (18)

The adoption of chatbots in the public sector has been slow, largely due to their limited application for straightforward information dissemination, which does not fully exploit their potential for enhancing public service delivery. Despite the increasing interest, these AI tools are often seen merely as aids for navigating information, falling short in meeting broader service delivery expectation. (19)

2.2. Emerging Trends in usage of Chatbots/Virtual assistants

With the emergence of Large Language Models (LLM) and Generative-AI platforms, platforms, some new capabilities emerge, along with potential challenges:

2.2.1. Capabilities

- **Context driven conversations:** Chatbots can now understand and respond to more complex conversations (20), including slang, idioms, and jokes (21).
- **Better emotional understanding:** Chatbots can now better understand the intent and emotions behind user input, thanks to advancements in NLP techniques like sentiment analysis and sarcasm detection. (22). Chatbots can now understand the emotional state of the person they're interacting with, which allows them to make better predictions and be more helpful. (20)
- **Industry-specific applications**: Chatbots are being used in a variety of industries, including healthcare, finance, and e-commerce. For example, healthcare chatbots can help users monitor their health, set medication reminders, and track symptoms. (23)
- **Enhanced Search Capabilities:** Large Language Models (*LLMs*) are being used to improve search functionality in government chatbots and virtual assistants. They can provide more relevant and comprehensive search results by understanding natural language queries and context. (24)
- **Efficient Policy Sharing:** RAG (*Retrieval Augmented Generation*) systems enable governments to efficiently share policies and regulations. By retrieving relevant information from vast databases, these systems can provide concise and accurate answers to user queries. (25)
- **Improved Accessibility:** Automatic Speech Recognition (*ASR*) technology allows users to interact with government chatbots and virtual assistants using voice commands. This improves accessibility for individuals with disabilities or those who prefer a hands-free experience. (26)

2.2.2. Challenges

- **Under-resourced languages:** Indian regional languages are regarded as under-resourced, meaning they possess fewer online materials, resources, and test data for developing text-to-speech or speech-to-text solutions compared to other languages. Of the entire population, barely 10% Indians use English to transact and most prefer regional languages, which have evolved over centuries. (11)

In his Book, “Measuring Linguistic Diversity in Indian Online Scenario”, Arul Selvan makes a case about how Indian languages have less content on the web despite each language having 10-100s of millions of speakers.

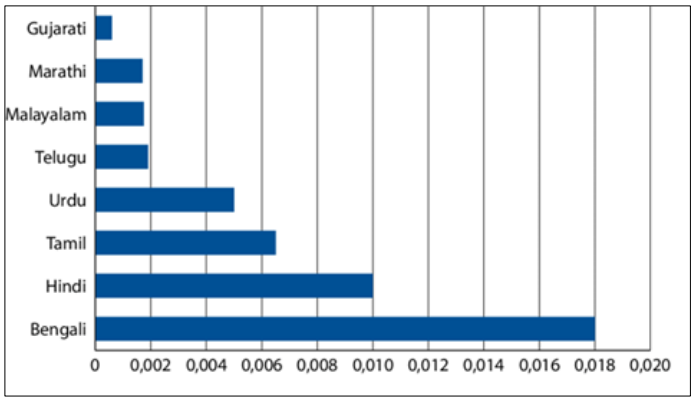

Figure 1 Ranking of Indian Languages based on web algorithm data (27)

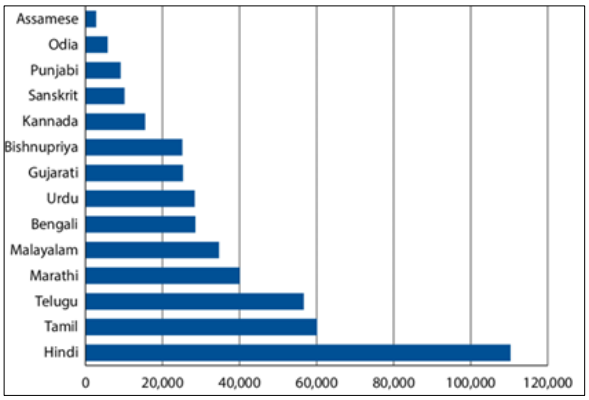

Figure 2 Ranking of Indian languages based on Wikipedia Articles (27)

As there is diversity in languages, language processing applications are a boon to the people for their day-to-day transaction. (11). This results in LLMs that scrape the web have less content in vernacular languages.

- **Hallucinations:** In Natural Language Generation, this is a serious issue as it can compromise performance and safety. Additionally, it raises privacy concerns, as models may unintentionally generate sensitive personal data from their training corpus, which is not present in the input. (28) . This is of two types:
  - Factuality Hallucination: This occurs when an LLM generates factually inaccurate content. (29)
  - Faithfulness Hallucination: These are cases where the model generates content that is inconsistent with the original source. (29)

2.3. Research Objectives

This research explores the efficacy of Retrieval-Augmented Generation (RAG) technologies and advancements in Large Language Models (LLMs). We extracted data on 2980 government programs through web scraping and utilized vector embedding to develop a solution capable of accepting both voice and text inputs in various vernacular languages as inputs to retrieve voice based responses. This solution translates the inputs into English, applies RAG techniques to fetch relevant information, and subsequently delivers the results back in the original language. The research is driven by the following objectives:

- Evaluate different LLM powered chat bots to cater to the information needs of target audience with special emphasis to the government schemes.
- Compare the performance of RAG architectures using different LLMs, embedding models, with and without Retrieval Augmentation.
- Identify measures using the following metrics and document performance for each architecture:
  - **Sentence Similarity Metrics:** Here we compare the generated text to the model's ground truth data or a reference corpus. For this metric, a higher score is desired.
  - **Semantic Uncertainty:** Even when the exact words are different, we will cluster model generations by meaning and then calculate semantic uncertainty. For this metric, a higher score is desired.
  - **Hallucination Metric:** This refers to a measure used to quantify the extent to which a language model generates text that is factually incorrect, irrelevant, or not grounded in reality

These are explained in detail in the Evaluation section.

3. Materials and Methods

This section has the following content:

- Data Ingestion Pipeline
- Conversational Chatbot Pipeline
- Application Repository
- Test setup

3.1. Data Ingestion Pipeline

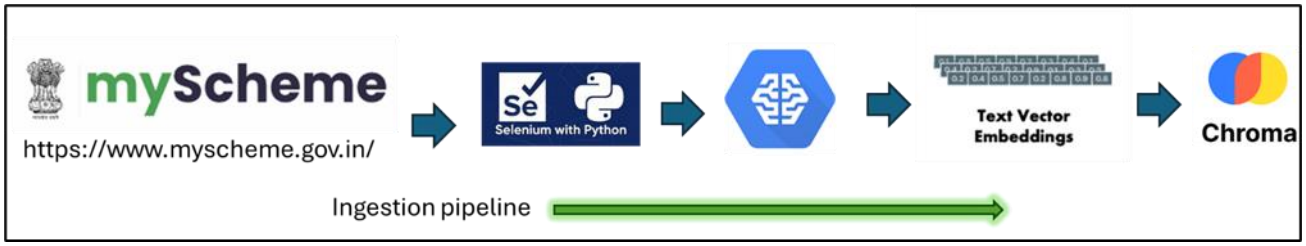

Figure 3 Ingestion Pipeline Architecture

- **Data Source:** Information is scraped from the mySchemе website using Selenium and Python.

Table 1 Data Ingested from government website

Attribute	Value
Data Source	https://www.myscheme.gov.in/search (30)
Date of Retrieval	Aug. 15, 2024
# of Records	2980 Schemes
Details Collected	Title, Details, Benefits, Eligibility Criteria, Application Process, Documents Required, Frequently asked questions and answers, Sources and References,
Ingestion Type	Batch Ingestion

- **Text Embeddings:** The scraped data is transformed into vector embeddings.
- **Chroma Database:** These embeddings are stored in Chroma, a vector database. Chroma DB is a fast, open-source vector store used for storing and retrieving vector embeddings.

3.1.1. Data Ingestion Pipeline Scraping Setup

Table 2 Tools and Libraries for Ingestion Pipeline

Tools	Version	Purpose
Python (31)	3.10.12	The programming language used for scripting and automation.
Selenium IDE (32)	4.24.0	A web automation tool used to interact with web elements dynamically. It helps in navigating through the web pages, handling timeouts, and extracting data from the website.
Selenium WebDriver (33)	4.24.0	The browser automation tool integrated with Selenium, used here with the Firefox browser.
Langchain's *JSONLoader* (34)	0.2.16	**Data Loading:** Loads scheme details from a JSON file and structures them for embedding generation.
Gemini models/embedding (35)	001	**Embedding Generation:** Converts textual data into numerical vectors that capture semantic meaning. Offers high-quality text embeddings, capturing semantic nuances effectively.

nomic-embed-text Embedding (36)	3.1	Converts textual data into numerical vectors that capture semantic meaning. Offers high-quality text embeddings, capturing semantic nuances effectively.
Chroma DB (37)	0.5.5	**Embedding Storage:** Indexes generated embeddings for efficient similarity searches. Provides efficient storage and retrieval of vector embeddings for fast similarity searches.

Code Functionality:

- **Initialization and Setup**: The scraper is set up with the MyScheme (30) website URL and uses Selenium WebDriver with Chrome/Firefox to handle dynamic content and JavaScript.
- **Collecting Scheme Links and Basic Info**: The scraper navigates the main page, waits for content to load, and extracts basic scheme details like the name, serial number, and detail page link. This information is stored in a list of dictionaries.
- **Extracting Detailed Information**: For each scheme, the scraper visits the detail page and extracts tags, descriptions, benefits, eligibility criteria, application processes, and required documents. Exception handling ensures smooth operation even with missing or unexpected content.
- **Compiling Data**: The detailed information is added to the original dictionaries. After collecting all data, the browser session is closed.
- **RAG Chain:** The system uses a Retrieval-Augmented Generation (RAG) Chain to retrieve relevant information and generate accurate responses.
- **Output**: The final output is stored in a JSON file, containing a structured list of dictionaries with comprehensive information about each scheme, ready for further processing or analysis.

3.1.2. Vector Embedding Logic

After scraping the MyScheme website and storing the scheme details in a JSON format, the next step involves converting this textual data into a form that can be efficiently searched and retrieved: vector embeddings. The vector embedding logic involves converting the textual data of government schemes into numerical vectors that capture the semantic meaning of the text. This process uses the **Gemini models/embedding-001**, and **nomic-embed-text** embeddings, to compare performance, as these are specifically designed for generating high-quality text embeddings, and Chroma DB for efficient storage and retrieval of these embeddings.

3.2. Conversational Chatbot Pipeline

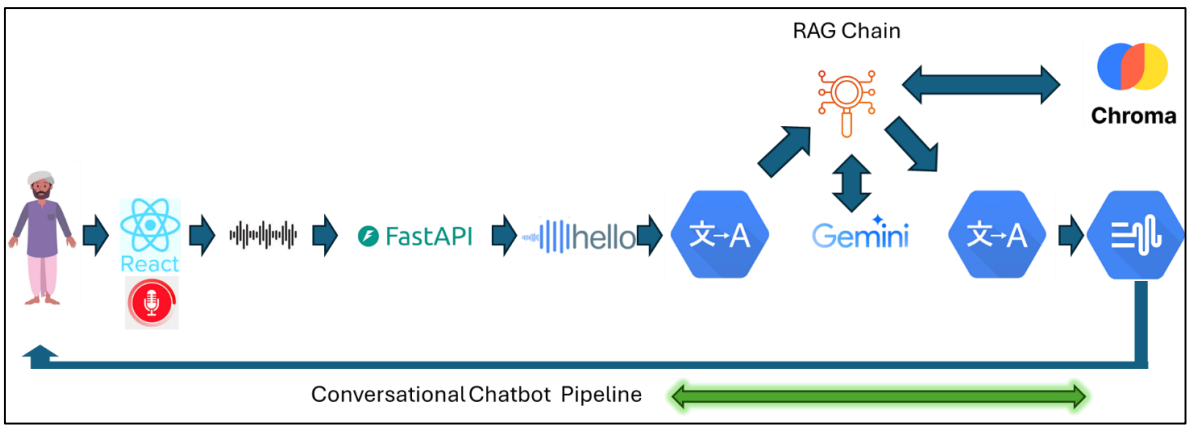

Figure 4 Retrieval Augmented Generation Pipeline

3.2.1. Conversational Chatbot Pipeline Setup

Tools and Libraries

Table 3 Tools and Libraries for Chatbot Pipeline

Tools	Version	Purpose
Front end (React) (38)	React v18.3.1	Users interact via a React-based interface.
FastAPI Backend (39)	0.114.0	Manages API interactions and processes user inputs.
Google Speech to Text API	V1	Handles conversion of speech to text (40)
Google Text-Speech API (41)		Handles conversion of text to speech
Google Translation API (42)		Handles multi-language processing.
Llama3 LLM Model (Llama3) (43)		A language model generating responses.
Gemini Pro LLM Model (35)		A language model generating responses.

Code Functionality

In essence, the system efficiently combines data retrieval and response generation to provide accurate and contextually appropriate chatbot interactions. The user is allowed to provide audio or text input using the React based user interface. It is converted to English language text using Google's translate and speech to text APIs. The English language text is fed into a semantic search and retrieval process.

3.2.2. Semantic Search and Information Retrieval:

Semantic Search and retrieval converts user queries into vectors and compares them with stored scheme embeddings to retrieve relevant results. The top 3 results are fed to LLM to generate response content in the user's input language. This response is shown and audio is played using text to speech apis. This approach enables accurate and relevant results, even with variations in input wording.

3.3. Application Repository

The application is hosted on GitHub, and the repository is available (44). The project setup and instructions for running the application are detailed in the repository's README file.

In the future, we plan to deploy the FastAPI (39) backend on Hugging Face, integrating Gradio to create an interactive interface. This will streamline the deployment process and make it easier to interact with the application.

3.4. Test Setup

In order to measure the performance of the LLM generated responses, we built a test setup. The setup was mainly used to query test questions, and validate the answer with desired answers. Since it was impossible to test a vast set of test cases manually, we built a test setup as follows

- We generated a corpus of 128 test questions from the scraped data using ChatGPT (45). We used the following prompt, and also attached the myschemes data as a .json file

I have attached a json file with a list of schemes with details such as scheme_name, details, eligibility, benefits, application_process, documents_required etc.

Generate 50 random questions for which the answers will be in the document. For each question, ensure it reflects only one scheme and related attribute. Give me the scheme name and attribute for each question, in the format provided below separated by new line:

Question: Your question

Scheme name: The scheme name

Attribute: One of the items in paranthesis (scheme_name, details, eligibility, benefits, application_process, documents_required)

Here is one example scheme in the document.

"scheme_name": "Garuda Scheme for Funeral Expense",

"details": "Details\nAndhra Pradesh Brahmin Welfare Corporation (ABC), a Government of Andhra Pradesh undertaking, is announcing the \"Garuda Scheme for Funeral Expenses\" (GS- FE) for the poor Brahmins in Andhra Pradesh. Under this scheme, financial assistance of \u20b910,000/- shall be provided to meet the Funeral Expenses of the deceased Brahmin, through the prescribed process. This scheme covers Andhra Pradesh all 13 districts, to meet the funeral expenses of the deceased Brahmi. ",

"benefits": "Benefits\nFinancial Assistance of \u20b910,000/- for funeral expenses.",

Example question you may generate:

Question: "I am from andhra pradesh, I am a brahmin. I need to get help on funeral expenses for my dad. Are there any government schemes and What are the eligibility criteria?"

scheme name: "Garuda Scheme for Funeral Expense"

Attribute: "eligibility"

Use different human tones in each question. Generate one question per scheme.

Figure 5 Prompt used to generate Test question Corpus

- Generate desired English response for each question – using SQL/text match and matching the scheme name and benefits
- For each type of embedding and each LLM, with and without RAG, generate responses from question corpus
- Save the responses in DB as follows: Embedding id, LLM id, Rag/norag (Y/N boolean), prompt used, result generated

3.4.1. Evaluation

Compare generated response with the original content and identify following metrics as described in the table below:

Table 4 Evaluation Metrics

Sentence Similarity Metrics:	Sentence similarity metrics are used to determine how similar two pieces of text are, typically on a scale that quantifies. Compare the generated text to the reference corpus. (46)
Semantic Uncertainty:	This refers to challenge of accurately interpreting the meaning of words, phrases, or entire texts due to inherent ambiguities, multiple possible interpretations, or limited context. (47)
Hallucination Metric	This refers to a measure used to quantify the extent to which a language model generates text that is factually incorrect, irrelevant, or not grounded in reality—essentially, "hallucinations." We compared the desired response against the actual response and asked ChatGPT (45) to score the hallucination metric for each response.

```
prompt = f"""
Desired: {desired}\n"
Actual: {predicted}\n"
  Evaluate how well the actual text matches the desired text in terms of meaning, coherence, and relevance.
Provide a score from 0 to 1, where 0 means the actual text is perfectly aligned with the desired text,
and 1 means the actual text is completely misaligned.
If the score is 0 or 1 print in float format like 0.0 or 1.0. Show me the score and the explanation in json format using keywords Score and Explanation.
Please provide the following information in strict JSON format without any additional text:
{{
 "Score": <the score value>,
 "Explanation": "<explanation text>"
}}
"""
```

Figure 6 Code generating prompt for computing Hallucination Score

4. Results and Discussions

4.1. Without RAG

It is observed that results with Retrieval perform better than results without Retrieval. We find that using the embedding model of the same family resulting in better quality responses.

Table 5 Performance Result when not using RAG

Score	LLM Models	Embedding Model	
		models/embedding-001	models/nomic-embed-text
Similarity Score	gemini-1.5-pro	0.317016829	0.317016829
	llama3	0.366839024	0.366839024
Uncertainty Score	gemini-1.5-pro	0.620319024	0.620319024
	llama3	0.629893415	0.629893415
Hallucination Score	gemini-1.5-pro	0.73902439	0.73902439
	llama3	0.707317073	0.707317073

4.2. With RAG

The experiment compares two embedding models and two LLM models across three metrics. The results show that for similarity scores, "models/embedding-001" performs better than "nomic-embed-text" across both LLMs. In terms of uncertainty, the two embedding models have comparable scores, while for hallucination, "nomic-embed-text" produces significantly higher values, indicating more hallucinations across both LLMs.

Table 6 Performance Results when using RAG

Score	LLM Models	Embedding Model	
		models/embedding-001	models/nomic-embed-text
Similarity Score	gemini-1.5-pro	0.463445122	0.221287317
	llama3	0.390919268	0.345552927
	gemini-1.5-pro	0.591999512	0.518791951

Uncertainty Score	llama3	0.619130488	0.623229024
Hallucination Score	gemini-1.5-pro	0.301219512	0.936585366
	llama3	0.618292683	0.863414634

Here are some sample results with RAG

Table 7 Results with RAG

Test #	Attribute	Configuration	Result
1	**Rag Used**	TRUE	The Garuda Scheme provides a financial assistance of â‚¹10,000/- to cover funeral expenses of the deceased. To be eligible, the applicant must be a close relative (spouse, son, daughter, parent, brother, or grandson) of the deceased Brahmin residing in Andhra Pradesh. They should also provide a self-declaration confirming their status as the legal heir. The applicant's annual family income should not exceed â‚¹75,000/-, and they should not have received similar benefits from other government schemes for the same purpose. The application process can be completed online through the Andhra Pradesh Brahmin Welfare Corporation website. Required documents include: - Passport-size photograph of the applicant - Aadhaar card of both the applicant and the deceased - White Ration Card/Income Certificate - Caste Certificate - Residence Certificate - Birth Certificate - Income Certificate of the applicant - Passbook of SB Account in Andhra Bank or SBI (in the name of the applicant) - Self-declaration (stating the applicant is the legal heir) - Death Certificate
	Embedding	models/embedding-001	
	LLM	gemini-1.5-pro	
	Prompt question	What type of financial assistance is available for funeral expenses under the Garuda Scheme?	
	Similarity	0.16756	
	Uncertainty	0.69189	
	Hallucination	0	
2	**Rag Used**	TRUE	The Indira Mahila Shakti Udyam Protsahan Yojana (IMSUPY) is a scheme by the Rajasthan government. Unfortunately, the provided information does not specify if loans are given for micro-enterprises. However, the scheme does offer financial assistance in the form of loans up to â‚¹50,00,000 for individual women and â‚¹1,00,00,000 for self-help groups. To be eligible, you must be a permanent resident of Rajasthan and 18 years old or above. Additionally, women's self-help groups or groups of these groups must be registered under a state government department. For clusters or federations of groups, registration under the Cooperative Act is necessary.
	Embedding	models/embedding-001	
	LLM	gemini-1.5-pro	
	Prompt question	How do I apply for a loan for micro-enterprises through the Indira Mahila Shakti Udyam Protsahan Yojana?	
	Similarity	0.38992	
	Uncertainty	0.59131	
	Hallucination	0.6	

			To apply, you need to register on the official portal using your Jan Aadhaar or Google account. After registering, you can log in and apply for the scheme by providing your details, project description, financial institution details, and uploading the required documents. Required documents include your Aadhaar card, address proof, bank account statement, passport-size photo, and mobile number.
3	**Rag Used**	TRUE	The "Immediate Relief Assistance" scheme under the "Welfare and Relief for Fishermen During Lean Seasons and Natural Calamities Scheme" provides financial assistance to families of fishermen who go missing in Puducherry. The family will receive â,¹ 1,00,000 in two installments. The first installment of â,¹ 50,000 is provided within 3 months of applying, and the remaining â,¹ 50,000 is deposited in a joint account after 9 months. This second installment will be released if there's no information about the missing person. If the fisherman returns, the amount will be recovered through an insurance bond. To be eligible, the applicant must be the legal heir of the missing fisherman, who should have been a resident of Puducherry, aged between 18-60, and a member of the Fishermen/Fisherwomen Co-operative Society. The fisherman must have gone missing while fishing and should not have been a beneficiary of the old age pension scheme. To apply, visit the Department of Fisheries and Fishermen Welfare or its Sub-Offices and obtain an application form. Submit the filled form with required documents like a photograph, residential certificate, age proof, a declaration stating the fisherman was not an old-age pension beneficiary, membership certificate, and other supporting documents like FIR and non-traceable certificate from the police, etc. Remember to apply within 30 days of the incident.
	Embedding	models/embedding-001	
	LLM	gemini-1.5-pro	
	Prompt question	How can a fishermanâ€™s family in Puducherry claim financial relief if the fisherman goes missing?	
	Similarity	0.68228	
	Uncertainty	0.53607	
	Hallucination	0.1	

5. Conclusion

Overall we find that leveraging Retrieval augmented generation using the same model embeddings substantially improves the quality of the output, its similarity to desired responses. This also reduces hallucinations.

We also notice that larger models perform better than smaller parameter models.

Governments can use these emerging technologies and build conversational chatbot solutions helping their citizen understand these schemes, policies better, resulting in overall better social outcomes.

As further research, we can include another dimension (quality of translations from and to indic languages) and see how it impacts overall performance. We could not do this research due to limitations in our testing infrastructure.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Ramla Khalidi. United Nations Development Programme. [Online].; 2022 [cited 2009 09 01. Available from: https://www.undp.org/vietnam/news/strengthening-social-protection-promote-sustainable-poverty-reduction.

[2] Srinivas Gurazada SKNAARaAFP. Government budget credibility and the impact of COVID-19. [Online].; 2022 [cited 2024 09 01. Available from: https://datatopics.worldbank.org/world-development-indicators/stories/government-budget-credibility-and-the-impact-of-covid-19.html.

[3] United Nations Statistics Division. The Sustainable Development Goals Report 2023: Special Edition. [Online].; 2023 [cited 2024 09 01. Available from: https://unstats.un.org/sdgs/report/2023/Goal-01/.

[4] Ahmad J. Schemes to Systems: The Future of Social Protection in India. [Online].; 2019 [cited 2024 09 01. Available from: https://www.worldbank.org/en/news/feature/2019/11/20/schemes-to-systems-future-social-protection-india.

[5] Asianet Newsable. This is why social welfare schemes fail in India. [Online]. [cited 2024 09 01. Available from: https://newsable.asianetnews.com/india/this-is-why-social-welfare-schemes-fail-in-india.

[6] Implementation National Statistics Office, National Statistics Office. DATA SNAPSHOT on Sustainable Development Goals National Indicator Framework Progress Report. New Delhi:; 2004.

[7] Singh RK. Failure Of Implementing Schemes And Policy In India: Who Is Responsible? [Online]. [cited 2024 09 01. Available from: https://www.legalserviceindia.com/legal/article-10121-failure-of-implementing-schemes-and-policy-in-india-who-is-responsible-.html.

[8] THE HANS INDIA. Why does government fail to implement a scheme successfully? [Online].; 2016 [cited 2024 09 01. Available from: https://www.thehansindia.com/posts/index/News-Analysis/2016-04-18/Why-does-government-fail-to-implement-a-scheme-successfully/222285?infinitescroll=1.

[9] Mahajan M. Why does the government fails when it comes to the "implementation "of the schemes? [Online].; 2018 [cited 2024 09 01. Available from: https://qr.ae/p2saCS.

[10] Dr. Pardeep Mittal AK. E-Governance - A challenge for India. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 2013; 2(3): 1196.

[11] B. S. Harish RKR. A comprehensive survey on Indian regional language processing. Springer Nature Switzerland AG 2020 SN Appl. Sci. 2, 1204 (2020). https://doi.org/10.1007/s42452-020-2983-x. June 2020; 2: https://rdcu.be/dSC6k.

[12] Namrata Prakash SoMGEHU. Challenges of E-Governance Implementation in Rural India: An Empirical. International Journal of Early Childhood Special Education (INT-JECSE). 2022; 14(5): 9027.

[13] Dubai Electricty and Water Authority. Dubai Electricty and Water Authority. [Online].; 2017 [cited 2024 09 01. Available from: https://www.dewa.gov.ae/en/about-us/media-publications/latest-news/2017/03/dewas-rammas-service-uses-artificial-intelligence-to-respond-to-customer-enquiries.

[14] Government of Mississippi. Get to know Missi. [Online].; 2017 [cited 2024 09 01. Available from: https://www.ms.gov/meet-missi.

[15] Patricia Kelly DGIa. Meet Alex – IP Australia's next-generation virtual assistant. [Online].; 2016 [cited 2024 09 01. Available from: https://www.wipo.int/wipo_magazine/en/2016/06/article_0006.html.

[16] Govtech Singapore. Activate public-facing chatbots and serve citizens better with VICA. [Online].; 2024 [cited 2024 09 01. Available from: https://www.tech.gov.sg/products-and-services/for-government-agencies/productivity-and-marketing/vica/.

[17] María E. Cortés-Cediel ASTICMPRB. Trends and challenges of e-government chatbots: Advances in exploring open government data and citizen participation content. [Online].; 2023 [cited 2024 09 01. Available from: https://www.sciencedirect.com/science/article/pii/S0740624X23000771#bb0005.

[18] Pablo Ramires Hernández DVCRVMM. Review on the Application of Artificial Intelligence-Based Chatbots in Public Administration. [Online].; 2023 [cited 2024 09 01. Available from: https://www.irma-international.org/chapter/review-on-the-application-of-artificial-intelligence-based-chatbots-in-public-administration/312625/.

[19] Tzuhao Chen MGHaME. The Adoption and Implementation of Artificial Intelligence Chatbots in Public Organizations: Evidence from U.S. State Governments. [Online].; 2023 [cited 2024 09 01. Available from: https://www.researchgate.net/publication/373879520_The_Adoption_and_Implementation_of_Artificial_Intelligence_Chatbots_in_Public_Organizations_Evidence_from_US_State_Governments.

[20] Adarsh. Evolution of Bots in 2024 And Future Trends. [Online].; 2024 [cited 2024 09 01. Available from: https://www.kommunicate.io/blog/evolution-of-chatbots-and-future-trends/.

[21] Dunham J. The Future of Conversational AI: Trends for 2024 and Beyond. [Online].; 2024 [cited 2024 09 01. Available from: https://www.chatbot.com/blog/future-of-conversational-ai/#:~:text=Innovations%20in%20AI%20models%20and%20interfaces&text=These%20advancements%20mean%20that%20chatbots,more%20natural%20and%20human%2Dlike.

[22] IBM. Conversational AI Examples, Applications & Use Cases. [Online].; 2024 [cited 2024 09 01. Available from: https://www.ibm.com/think/topics/conversational-ai-use-cases#:~:text=As%20conversational%20AI%20continues%20to,more%20natural%20and%20engaging%20interactions.

[23] Denis Sinelnikov CoMCAWFSDMA. Advancements in AI Chatbots: Transitioning from Basic Scripts to Intelligent Conversations. [Online].; 2023 [cited 2024 09 01. Available from: https://www.linkedin.com/pulse/advancements-ai-chatbots-transitioning-from-basic-denis-sinelnikov-qgbje/.

[24] Shervin Minaee TMNNMCRSXAJG. Large Language Models: A Survey. [Online].; 2024 [cited 2024 09 01. Available from: https://arxiv.org/abs/2402.06196.

[25] Huayang Li YSDCYWLL. A Survey on Retrieval-Augmented Text Generation. [Online].; 2022 [cited 2024 09 01. Available from: https://arxiv.org/abs/2202.01110.

[26] Gupta A, Kumar R, Kumar Y. An Automatic Speech Recognition System: A systematic review and Future directions. In 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 1492-1496, doi: 10.1109/ICAC3N56670.2022.10074009.; 2022; Greater Noida, India.

[27] Selvan A. Measuring Linguistic Diversity in Indian Online Scenario Gothenburg: The International Clearinghouse on Children, Youth and Media Nordicom University of Gothenburg; 2015.

[28] ZIWEI JI NLRFTYDSYXEI. Survey of Hallucination in Natural Language Generation. ACM Computing Survey. 2022; 1(1).

[29] Shah D. The Beginner's Guide to Hallucinations in Large Language Models. [Online].; 2023 [cited 2024 09 01. Available from: https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models#:~:text=Hallucinations%20are%20a%20critical%20obstacle,and%20the%20models'%20interpretative%20limits.&text=of%20LLM%20Hallucination-,Mitigating%20Hallucinations%20in%20Large%20Lan.

[30] Government of India. myScheme. [Online].; 2024 [cited 2024 09 01. Available from: https://www.myscheme.gov.in/.

[31] Ver: 3.10.12.

[32] Conservancy SF. Selenium Webdriver. 2024..

[33] Software Freedom Conservancy..

[34] LangChain. Langchain Documentloader..

[35] Google. Gemini Embeddings API..

[36] Nomic.ai. nomic-embed-text. 2024 2024 09..

[37] Apache. Apache Chroma DB. 2024..

[38] React. React. [Online]. [cited 2024 09 01. Available from: https://react.dev/.

[39] Github - Fast API. [Online]. [cited 2024 09 01. Available from: https://github.com/fastapi/fastapi.

[40] Google, Inc. Speech-to-Text AI: speech recognition and transcription | Google Cloud. [Online]. [cited 2024 09 01. Available from: https://cloud.google.com/speech-to-text/?utm_source=google&utm_medium=cpc&utm_campaign=japac-IN-all-en-dr-BKWS-all-hv-trial-PHR-dr-1605216&utm_content=text-ad-none-none-DEV_c-CRE_634266628477-ADGP_Hybrid+%7C+BKWS+-+BRO+%7C+Txt+-AI+%26+ML-Speech+to+Text-go.

[41] Google, Inc. Text-to-Speech AI: Lifelike Speech Synthesis | Google Cloud. [Online]. [cited 2024 09 01. Available from: https://cloud.google.com/text-to-speech?hl=en.

[42] Google, Inc. Cloud Translation API | Google Cloud. [Online]. [cited 2024 09 01. Available from: https://cloud.google.com/translate/docs/reference/rest.

[43] Meta, Inc. Llama 3.1. 2024..

[44] GitHub - Vaishnav2804/UdhaviBot: UdhaviBot is a chatbot designed to assist and empower underserved individuals by providing information on Indian government policies. [Online].; 2024 [cited 2024 09 01. Available from: https://github.com/Vaishnav2804/UdhaviBot.

[45] OpenAI, Inc. Chatgpt Response. Prompt: I have attached a json file with a list of schemes with details such as scheme_name, details, eligibility, benefits, application_process, documents_required etc. Generate me 50 random questions for which the answers will be in the document. For ea.

[46] deepchecks. deepchecks. [Online].; 2026 [cited 2024 09 01. Available from: https://deepchecks.com/llm-hallucination-detection-and-mitigation-best-techniques/#:~:text=a%20similarity%20score.-,Detection%20Method,text%20may%20be%20a%20hallucination.

[47] Sebastian F, Jannik K, Lorenz , Yarin G. Detecting hallucinations in large language models using semantic entropy. Nature. 2024 Jun 19.

[48] Dice Staff. AI Jobs Demand Analysis: Current Trends and Future Outlook in Artificial Intelligence Careers. [Online].; 2024 [cited 2024 09 01. Available from: https://www.dice.com/career-advice/ai-jobs-demand-analysis-current-trends-and-future-outlook#:~:text=According%20to%20Grandview%20Research%2C%20global,reach%20%241.8%20billion%20by%202030.

[49] Google, Inc. Google Chrome – The Fast and Secure Web Browser Built to be Yours. [Online]. [cited 2024 09 01. Available from: https://www.google.com/chrome/.