AUTUMN INTERNSHIP PROJECT REPORT

Design and Development of a RAG-Based AI Assistant for IDEAS-TIH Using Open-Source Large Language Models

Kingsuk Ghosh
2nd Year, B.Tech. in CSE
Netaji Subhash Engineering College, Kolkata

Under the Mentorship of:
Agnimitra Biswas

Internship Duration:
26 August 2025 – 15 November 2025

Submitted to:
IDEAS-TIH, Indian Statistical Institute (ISI), Kolkata

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to **IDEAS-TIH, Indian Statistical Institute (ISI), Kolkata**, for providing me with the valuable opportunity to participate in the **Autumn Internship 2025**. This internship has been immensely beneficial in enhancing my knowledge, technical skills, and understanding of real-world applications of Artificial Intelligence.

I am deeply grateful to **my mentor, Mr. Agnimitra Biswas**, for his continuous guidance, encouragement, and insightful feedback throughout the internship period. His expertise and support were crucial for the successful completion of this project.

I also convey my heartfelt thanks to the entire **IDEAS-TIH team** for their resources, cooperation, and support during the internship. Finally, I extend my gratitude to my college, **Netaji Subhash Engineering College**, for enabling me to participate in such a valuable learning opportunity.

TABLE OF CONTENTS

Page

1. Introduction……………………………………………………4

2. Objectives………………………………………………………...5

3. Methodology…………………………………………………….6-9

4. Tools & Technologies Used……………………………………..10

5. Implementation Details………………………………………….11

6. Challenges Faced & Solutions…………………………………..11

7. Results / Outcomes……………………………………………12-13

8. Conclusion………………………………………………………14

9. References……………………………………………………....14

INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) has transformed how information is accessed, processed, and delivered. Modern organizations increasingly depend on AI assistants and chatbots to provide quick, accurate, and user-friendly access to domain-specific information.

IDEAS-TIH (Institute of Data Engineering, Analytics and Science Foundation), established under NM-ICPS at ISI Kolkata, has a diverse set of programs, activities, research initiatives, internships, and events. However, accessing this information manually from the website can be time-consuming.

To solve this challenge, the goal of this project was to design and develop an intelligent **AI-powered chatbot** capable of retrieving information about IDEAS-TIH directly from its website using **Retrieval-Augmented Generation (RAG)**. The system would scrape official webpages, process them into a searchable knowledge base, and allow users to ask natural-language questions and receive accurate, context-based answers.

This solution combines **web scraping, text preprocessing, vector embeddings, FAISS similarity search, and an open-source large language model (Qwen2.5)** to build a completely free, offline-capable, efficient, and domain-specific AI assistant.

This project demonstrates the ability to integrate NLP, machine learning, and information retrieval techniques into a practical application.

OBJECTIVES

The primary objectives of the project were:

1. Build an AI Assistant for IDEAS-TIH

Develop a chatbot capable of answering queries related to IDEAS-TIH using website data.

2. Implement a Full RAG (Retrieval-Augmented Generation) Pipeline

Combine retrieval (FAISS search) + generation (Qwen2.5 LLM) for accurate, grounded answers.

3. Use Free and Open-Source Tools

Ensure the system runs without paid APIs or proprietary restrictions.

4. Automate Website Data Extraction

Scrape multiple webpages from IDEAS-TIH to create a structured knowledge base.

5. Enable Terminal-Based Interaction

Build a fully operational chatbot interface inside Google Colab.

6. Prepare the System for Future Web Integration

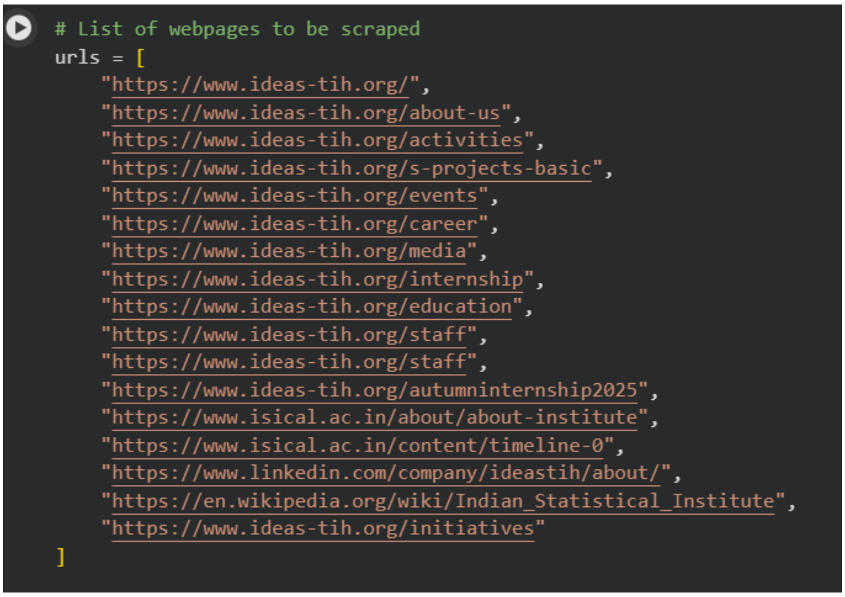
Ensure backend modularity for future frontend (Wix) integration.

METHODOLOGY

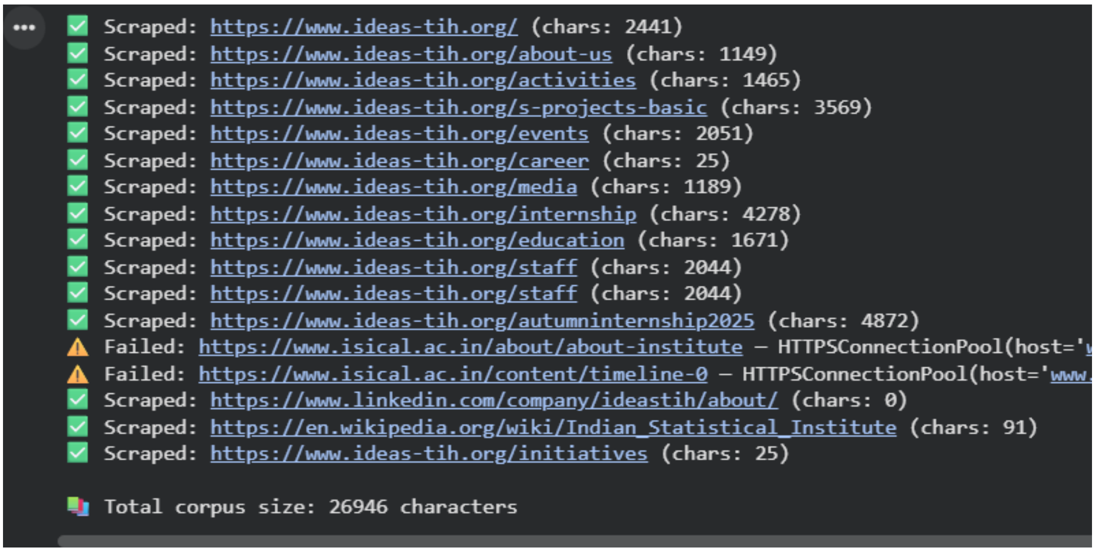
The project followed a systematic process:

STEP 1 — Web Scraping

- Using Python requests + BeautifulSoup
- Extracted text from multiple IDEAS-TIH webpages
- Combined into a unified corpus



```
# List of webpages to be scraped
urls = [
    "https://www.ideas-tih.org/",
    "https://www.ideas-tih.org/about-us",
    "https://www.ideas-tih.org/activities",
    "https://www.ideas-tih.org/s-projects-basic",
    "https://www.ideas-tih.org/events",
    "https://www.ideas-tih.org/career",
    "https://www.ideas-tih.org/media",
    "https://www.ideas-tih.org/internship",
    "https://www.ideas-tih.org/education",
    "https://www.ideas-tih.org/staff",
    "https://www.ideas-tih.org/staff",
    "https://www.ideas-tih.org/autumninternship2025",
    "https://www.isical.ac.in/about/about-institute",
    "https://www.isical.ac.in/content/timeline-0",
    "https://www.linkedin.com/company/ideastih/about/",
    "https://en.wikipedia.org/wiki/Indian_Statistical_Institute",
    "https://www.ideas-tih.org/initiatives"
]
```



```
✅ Scraped: https://www.ideas-tih.org/ (chars: 2441)
✅ Scraped: https://www.ideas-tih.org/about-us (chars: 1149)
✅ Scraped: https://www.ideas-tih.org/activities (chars: 1465)
✅ Scraped: https://www.ideas-tih.org/s-projects-basic (chars: 3569)
✅ Scraped: https://www.ideas-tih.org/events (chars: 2051)
✅ Scraped: https://www.ideas-tih.org/career (chars: 25)
✅ Scraped: https://www.ideas-tih.org/media (chars: 1189)
✅ Scraped: https://www.ideas-tih.org/internship (chars: 4278)
✅ Scraped: https://www.ideas-tih.org/education (chars: 1671)
✅ Scraped: https://www.ideas-tih.org/staff (chars: 2044)
✅ Scraped: https://www.ideas-tih.org/staff (chars: 2044)
✅ Scraped: https://www.ideas-tih.org/autumninternship2025 (chars: 4872)
⚠ Failed: https://www.isical.ac.in/about/about-institute — HTTPSConnectionPool(host='w
⚠ Failed: https://www.isical.ac.in/content/timeline-0 — HTTPSConnectionPool(host='www.
✅ Scraped: https://www.linkedin.com/company/ideastih/about/ (chars: 0)
✅ Scraped: https://en.wikipedia.org/wiki/Indian_Statistical_Institute (chars: 91)
✅ Scraped: https://www.ideas-tih.org/initiatives (chars: 25)

📚 Total corpus size: 26946 characters
```

STEP 2 — Text Preprocessing

- Cleaned raw HTML text
- Chunked into smaller segments (600 words each)

STEP 3 — Vector Embedding

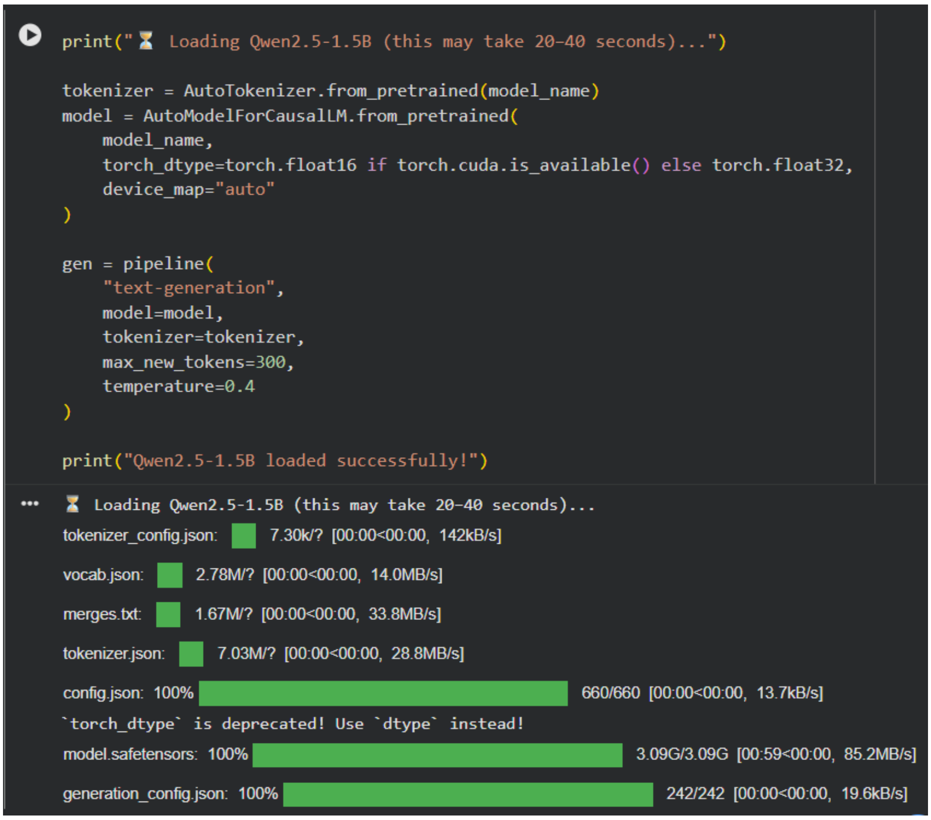
- Used all-MiniLM-L6-v2 SentenceTransformer to convert text into 384-dimensional embeddings
- Free, fast, local, ideal for FAISS search

STEP 4 — Building a FAISS Vector Index

- Indexed embeddings into FAISS IndexFlatL2
- Enabled fast similarity search (top-k chunks)

STEP 5 — Loading the Qwen2.5 LLM

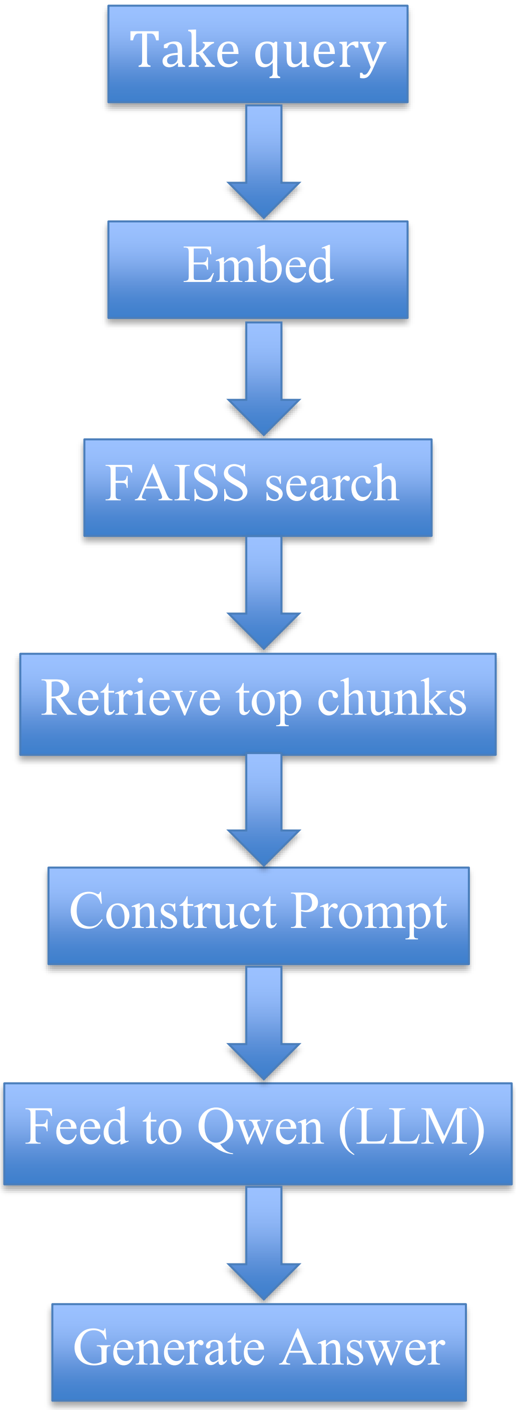
- Loaded Qwen2.5-1.5B-Instruct (open-source, fine-tuned for instruction following)
- Deployed via HuggingFace transformers locally in Colab



```
print("⏳ Loading Qwen2.5-1.5B (this may take 20-40 seconds)...")

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    torch_dtype=torch.float16 if torch.cuda.is_available() else torch.float32,
    device_map="auto"
)

gen = pipeline(
    "text-generation",
    model=model,
    tokenizer=tokenizer,
    max_new_tokens=300,
    temperature=0.4
)

print("Qwen2.5-1.5B loaded successfully!")
```

```
⏳ Loading Qwen2.5-1.5B (this may take 20-40 seconds)...
tokenizer_config.json: 7.30k/? [00:00<00:00, 142kB/s]
vocab.json: 2.78M/? [00:00<00:00, 14.0MB/s]
merges.txt: 1.67M/? [00:00<00:00, 33.8MB/s]
tokenizer.json: 7.03M/? [00:00<00:00, 28.8MB/s]
config.json: 100% 660/660 [00:00<00:00, 13.7kB/s]
`torch_dtype` is deprecated! Use `dtype` instead!
model.safetensors: 100% 3.09G/3.09G [00:59<00:00, 85.2MB/s]
generation_config.json: 100% 242/242 [00:00<00:00, 19.6kB/s]
```

STEP 6 — RAG Pipeline Integration



Take query

Embed

FAISS search

Retrieve top chunks

Construct Prompt

Feed to Qwen (LLM)

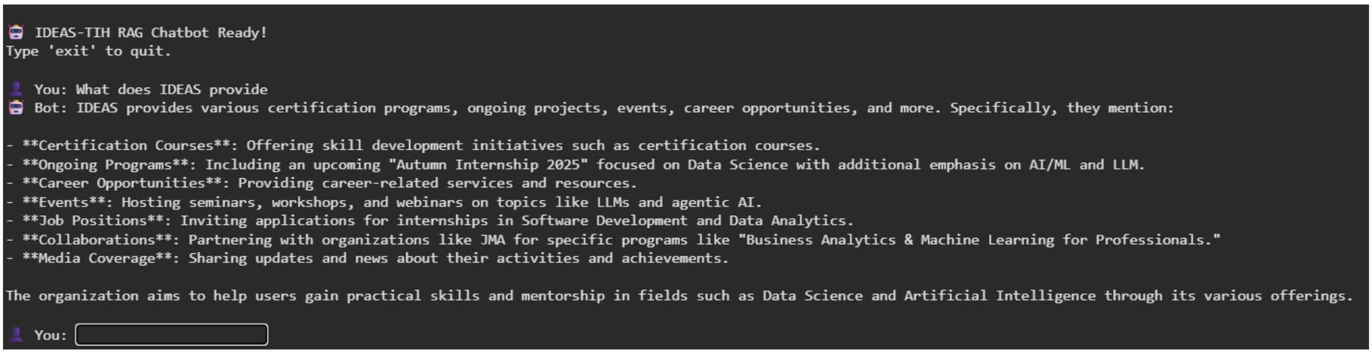
Generate Answer

STEP 7 — Terminal Chatbot Interface

- Created interactive loop
- User enters question → bot replies using RAG

STEP 8 — Testing and Validation

- Verified responses for accuracy and grounding



```
🤖 IDEAS-TIH RAG Chatbot Ready!
Type 'exit' to quit.

👤 You: What does IDEAS provide
🤖 Bot: IDEAS provides various certification programs, ongoing projects, events, career opportunities, and more. Specifically, they mention:

- **Certification Courses**: Offering skill development initiatives such as certification courses.
- **Ongoing Programs**: Including an upcoming "Autumn Internship 2025" focused on Data Science with additional emphasis on AI/ML and LLM.
- **Career Opportunities**: Providing career-related services and resources.
- **Events**: Hosting seminars, workshops, and webinars on topics like LLMs and agentic AI.
- **Job Positions**: Inviting applications for internships in Software Development and Data Analytics.
- **Collaborations**: Partnering with organizations like JMA for specific programs like "Business Analytics & Machine Learning for Professionals."
- **Media Coverage**: Sharing updates and news about their activities and achievements.

The organization aims to help users gain practical skills and mentorship in fields such as Data Science and Artificial Intelligence through its various offerings.

👤 You:
```

TOOLS & TECHNOLOGIES USED

1. Python

Primary programming language for all development.

2. Google Colab

- Free GPU runtime
- Easy for running large models
- Ideal environment for LLM experimentation

3. BeautifulSoup4

For web scraping and extracting meaningful text from HTML.

4. SentenceTransformers

- all-MiniLM-L6-v2 used for embeddings
- Free and efficient for semantic search

5. FAISS (Facebook AI Similarity Search)

- State-of-the-art vector database
- Enables fast top-k retrieval

6. HuggingFace Transformers

Used to load and run the Qwen2.5 LLM locally.

7. Qwen/Qwen2.5-1.5B-Instruct

- Open-source LLM
- No API needed
- Efficient and accurate

IMPLEMENTATION DETAILS

The system scrapes multiple URL pages from IDEAS-TIH, chunks the text, embeds them with MiniLM, indexes via FAISS, and generates accurate responses using Qwen2.5 in a RAG loop.

CHALLENGES FACED & SOLUTIONS

1. Gemini and HuggingFace API Restrictions

- Limited free tier
- Authentication failures
- Rate limit issues
  Solution: Switched to fully open-source Qwen2.5

2. Embedding Quota Limitations

- Gemini embeddings blocked in free tier
  Solution: Used all-MiniLM-L6-v2 SentenceTransformer

3. Model Loading Delays

- Large models load slowly on CPU
  Solution: Used Google Colab GPU + small 1.5B model

4. Preventing Hallucination

Solution: A strict RAG prompt was implemented

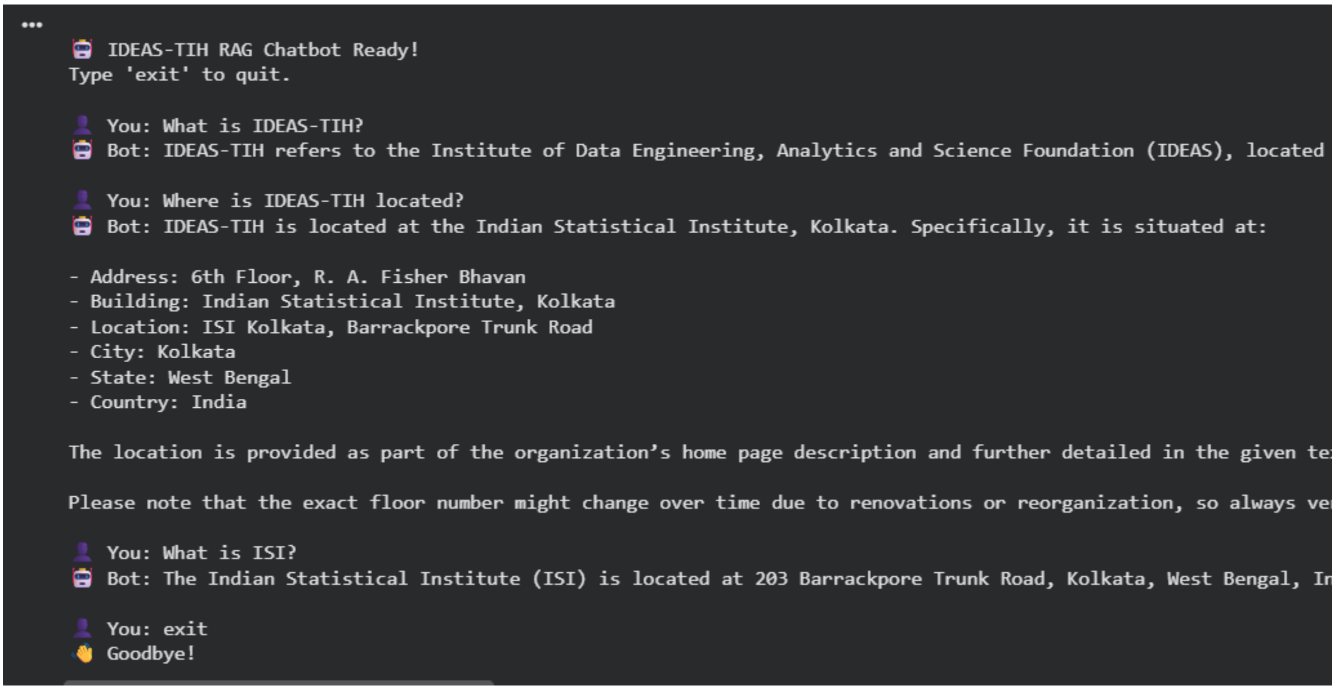
RESULTS / OUTCOME

Working RAG chatbot for IDEAS-TIH

- Answers domain-specific questions like:
    - "What is IDEAS-TIH?"
    - "Where is IDEAS located?"
    - "What programs does IDEAS organize?"

Accurate, context-grounded responses

- Bot answers using retrieved chunks ONLY
- Prevents hallucination
- Follows professional tone



```
🤖 IDEAS-TIH RAG Chatbot Ready!
Type 'exit' to quit.

👤 You: What is IDEAS-TIH?
🤖 Bot: IDEAS-TIH refers to the Institute of Data Engineering, Analytics and Science Foundation (IDEAS), located

👤 You: Where is IDEAS-TIH located?
🤖 Bot: IDEAS-TIH is located at the Indian Statistical Institute, Kolkata. Specifically, it is situated at:

- Address: 6th Floor, R. A. Fisher Bhavan
- Building: Indian Statistical Institute, Kolkata
- Location: ISI Kolkata, Barrackpore Trunk Road
- City: Kolkata
- State: West Bengal
- Country: India

The location is provided as part of the organization's home page description and further detailed in the given te

Please note that the exact floor number might change over time due to renovations or reorganization, so always ve

👤 You: What is ISI?
🤖 Bot: The Indian Statistical Institute (ISI) is located at 203 Barrackpore Trunk Road, Kolkata, West Bengal, In

👤 You: exit
👋 Goodbye!
```

Fully offline and free

- Runs entirely on open-source models
- No API restrictions
- Suitable for deployment and demonstration

Modular Codebase

- Ready for backend integration
- Easily extendable to Wix frontend chatbot

CONCLUSION

The project successfully achieved its core objectives. A complete Retrieval-Augmented Generation (RAG) chatbot was built for IDEAS-TIH using entirely free and open-source technologies. The system performs domain-specific information retrieval with high accuracy and conversational fluency. It demonstrates practical application of NLP concepts such as embeddings, similarity search, and generative modeling.

By avoiding proprietary APIs and relying solely on open-source tools, the solution is future-proof, customizable, and suitable for further deployment on websites like Wix. This project has strengthened my understanding of AI systems, LLMs, vector databases, and real-world chatbot design.

REFERENCES

1. Qwen2.5 LLM – https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct

2. SentenceTransformers – https://www.sbert.net/

3. FAISS – https://faiss.ai/

4. BeautifulSoup – https://www.crummy.com/software/BeautifulSoup/

5. Transformers – https://huggingface.co/docs/transformers

6. IDEAS-TIH – https://www.ideas-tih.org/