

AI-APPLICATION SIMPLIFYING STUDY ABROAD PROCESSES IN NEPAL

**SUBMITTED BY :
JIVAN ACHARYA**

Objective

- Provides personalized guidance to students about universities, courses, and scholarships, with aid in document preparation
- Reduces dependency on expensive consultancy services by offering free or affordable AI-driven solutions.

Features

- Chat based interface for searching and querying university their courses and scholarship schemes.
- Agentic Professor Search feature.
- Document preparation tool (SOP review).

Feasibility Study

- Focusing on a single country primarily for Graduate Studies
- Data collection through ranking websites, Kaggle datasets, university websites
- Open source LLMs such as Deepseek, frameworks such as Langchain, agentic frameworks such as CrewAI, SmolaAgents
- Retrieval Augmented Generation approach for updated and relevant information retrieval

Functional Requirements - [1]

1. User Authentication & Profile Management

- The system should allow users to sign up and log in.
- Users should be able to track chat history for personalized recommendations.

2. Chatbot Interface

- Should have a user-friendly chat interface where users can interact with three key modules:
 - General Queries (University Search) – RAG-Based
 - Professor Search – Agentic Workflow
 - SOP Review – LLM-Based
- It should allow follow-up questions within the same session.

Functional Requirements - [2]

3. University Search (RAG-Based)

- The user should be able to enter queries related to universities, such as: "Best universities in California for AI research", "Universities with low tuition fees for international students"
- The system should use similarity search in a vector database to retrieve relevant university information.
- The retrieved results should be passed to the LLM, which will generate a well-structured response.
- The retrieved result should be checked if it aligns with the response or not, if not it should regenerate

Functional Requirements - [3]

4. Professor Search (Agentic Workflow)

- Users should be able to search for professors based on criteria like: Research Interests, University Affiliation, Number of Publications, etc
- The system should retrieve professor profiles information from the existing dataset or the web
- Retrieve professor profiles based on relevance.
- Summarize each professor's research focus and publications.

Functional Requirements - [4]

5. SOP Review (LLM-Based)

- Users should be able to upload or paste their Statement of Purpose (SOP) for review.
- The LLM should analyze the SOP for: Grammar & Style, Content Structure, Clarity & Coherence, Strength of Research Interests
- The system should provide feedback and suggest improvements.
- The chatbot should provide examples of well-written SOPs for reference.

Functional Requirements - [5]

6. Chat History & User Data Management

- The system should store chat history for each user
- Users should be able to delete or download their past chats.
- The chatbot should maintain session continuity, allowing users to resume previous interactions.
- The chatbot should provide a REST API for external integrations.

Non - Functional Requirements

1. Performance

- The chatbot should respond within 2 to 5 seconds for most queries.
- The system should support at least 100 concurrent users without performance degradation.

2. Scalability

- The system should be designed modularly to support future expansions.
- The vector database should handle large embeddings efficiently.

3. Availability & Reliability

4. Security

- User credentials should be stored using hashed passwords (bcrypt or Argon2).

5. User Experience (UX)

- The UI should be responsive and mobile-friendly.

Data Collection

- QSRankings.com Data for Top Universities Data
- CSRankings.com for Professor/ Faculty Data
- TimesHigherEducation.com data for University Details

Data Preprocessing

- Data Cleaning (missing values , features extraction and filtering)
- Merging Datasets (Univeristy and Faculty information)
- Data Expansion (Json to document type using LLMs)
- Data Splitting and Chunking
- VectorStore Creation

Data Snapshots - [1]

```
[  
 {  
   "rank": "2",  
   "name": "Massachusetts Institute of Technology",  
   "aliases": "Massachusetts Institute of Technology",  
   "scores_citations": 99.7,  
   "scores_research": 96.0,  
   "scores_teaching": 99.2,  
   "stats_number_students": "11,836",  
   "stats_pc_intl_students": "33%",  
   "subjects_offered": "Biological Sciences, Electrical & Electronic Engineering, Language & Literature, Mathematics & Statistics, Physics",  
   "Location": "Cambridge, United States",  
   "Description": "The Massachusetts Institute of Technology (MIT) is an independent, privately funded research university located in Cambridge, Massachusetts.",  
   "uni_url": "https://www.mit.edu/"  
 },  
 {  
   "rank": "3",  
   "name": "Harvard University",  
   "aliases": "Harvard University",  
   "scores_citations": 99.3,  
   "scores_research": 99.9,  
   "scores_teaching": 97.3,  
   "stats_number_students": "22,584",  
   "stats_pc_intl_students": "25%",  
   "subjects_offered": "Archaeology, Mathematics & Statistics, Art, Performing Arts & Dance, English, History, International Relations, Italian, Latin American Studies, Linguistics, Near Eastern Languages & Civilizations, Philosophy, Political Science, Psychology, Religious Studies, Sociology, Spanish, Theater & Drama",  
   "Location": "Cambridge, United States",  
   "Description": "Dating back to 1636, Harvard University is the oldest university in the United States.",  
   "uni_url": "https://www.harvard.edu/"  
 }]
```

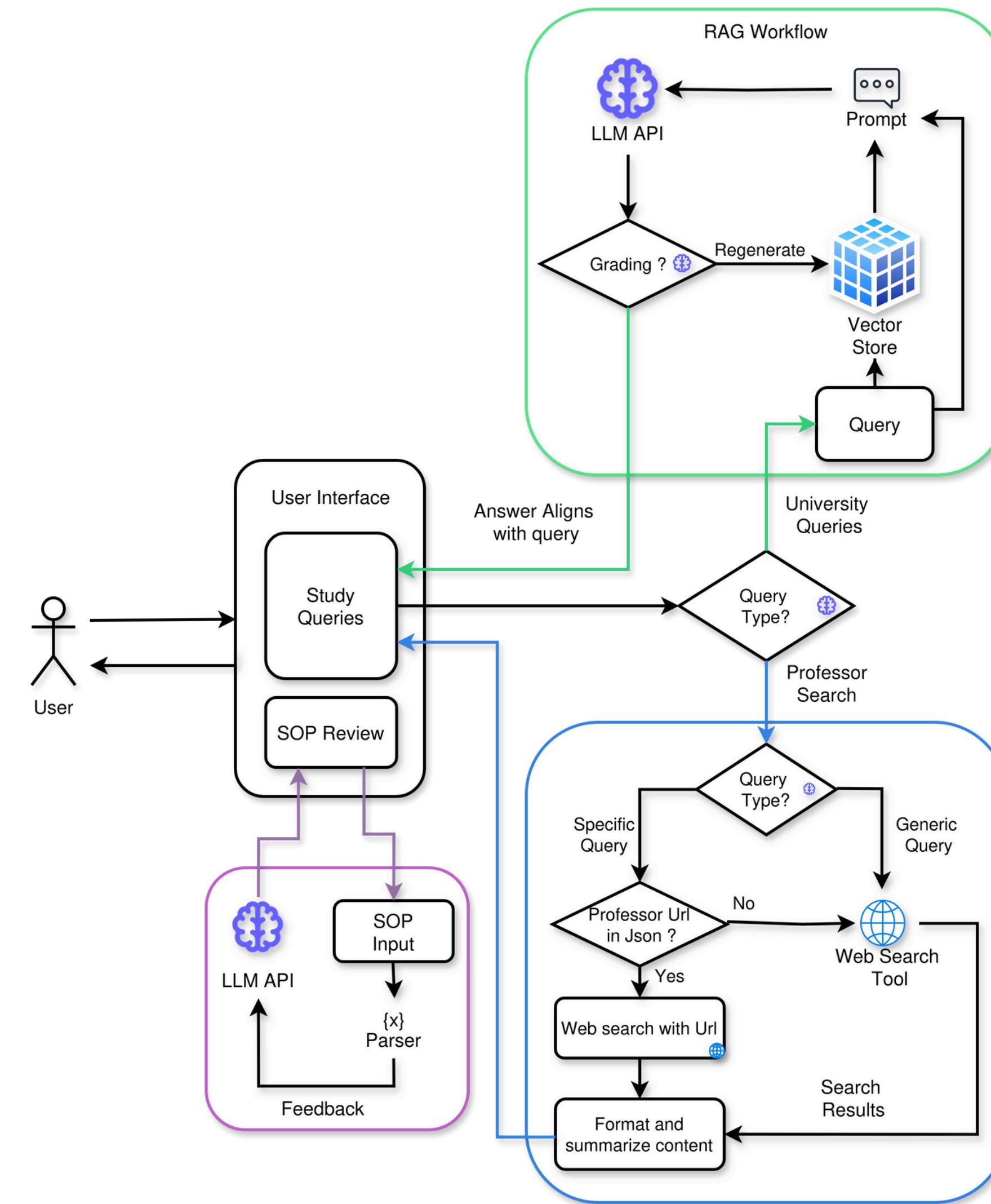
```
{  
    "Carnegie Mellon University": [ ...  
    ],  
    "Univ. of Illinois at Urbana-Champaign": [ ...  
    ],  
    "Univ. of California - San Diego": [ ...  
    ],  
    "Georgia Institute of Technology": [  
        [  
            "A. S. Nemirovskii 0001",  
            "https://www.isye.gatech.edu/users/arkadi-nemirovski"  
        ],  
        [  
            "Ada Gavrilovska",  
            "http://www.cc.gatech.edu/home/ada"  
        ],  
        [  
            "Ahmed Saeed 0001",  
            "https://www.cc.gatech.edu/~amsmti3"  
        ],  
        [  
            "John Doe 0001",  
            "https://www.cc.gatech.edu/~johndoe"  
        ]  
    ]  
}
```

Data Snapshots - [2]

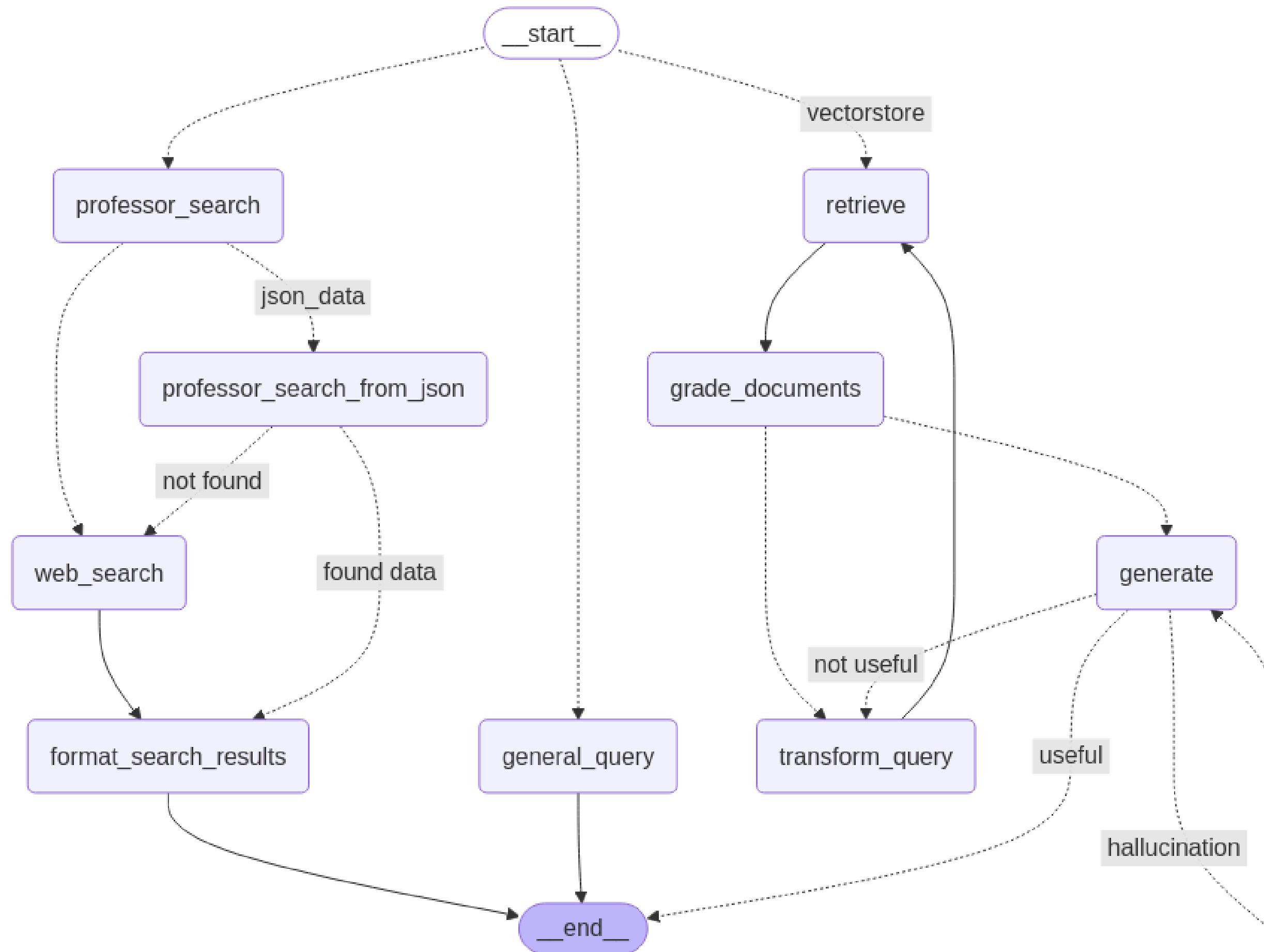
Massachusetts Institute of Technology
Rank: 2 (National), 1 (Engineering)
Location: Cambridge, MA, United States
Aliases: MIT
Number of Students: 11,836 (4,657 graduate)
Percentage of International Students: 33%
Subjects Offered: Biological Sciences, Electrical & Electronic Engineering, Computer Science, Business & Management
Graduate Programs: Master's in Computer Science (2 years, on-campus), PhD in Electrical Engineering (4-6 years, on-campus)
Admissions: Min GPA 3.5, GRE optional, TOEFL 90+, 3 LORs, SOP, resume; Deadline: Dec 15; Fee: \$75; Acceptance Rate: 35%
Tuition: \$57,590/year (2024-25)
Funding: Need-based aid, fellowships (e.g., MIT Presidential Fellowship), TA/RA positions (\$30K-\$40K/year)
Housing: On-campus options (\$10K-\$15K/year), off-campus avg \$1,500/month
Career Outcomes: 92% job placement within 6 months, avg salary \$110K (engineering)
Visa: OPT up to 36 months for STEM
Website: <https://www.mit.edu/>, <https://gradadmissions.mit.edu/>

Harvard University
Rank: 3 (National), 1 (Law), 1 (Business)
Location: Cambridge, MA, United States
Aliases: Harvard
Number of Students: 30,631 (21,263 graduate)
Percentage of International Students: 23%
Subjects Offered: Law, Business & Management, Medicine, Engineering, Public Health, Education, Social Sciences
Graduate Programs: Master of Laws (LL.M., 1 year, on-campus), Master of Business Administration (MBA, 2 years, on-campus)
Admissions: Min GPA 3.5, GRE optional (varies by program), TOEFL 100+ or IELTS 7.0+, 3 LORs, SOP, resume; Deadline: Dec 15; Fee: \$1,000
Tuition: \$50,654/year (2024-25)
Funding: Need-based aid, fellowships (e.g., Harvard Griffin GSAS Fellowship), TA/RA positions (\$40K-\$50K/year)
Housing: On-campus options (\$1,500-\$2,500/month), off-campus avg \$2,500/month
Career Outcomes: 86% job placement within 6 months, avg salary \$91K (overall), \$175K (MBA)
Visa: OPT up to 36 months for STEM
Website: <https://www.harvard.edu/>, <https://gsas.harvard.edu/admissions>

System Workflow



LangGraph workflow



Testing of chat history and context retrieval

Chat Sessions

New Chat

Session: 08094d23...

University Queries Chatbot

Chat Session: 08094d23...

List universities in Texas

There are two universities mentioned in the retrieved information:

1. Texas State University
2. University of Texas at Austin

What are the subjects offered in the first one?

At Texas State University, the following subjects are offered:

1. Geography
2. Languages, Literature & Linguistics
3. Biological Sciences

Routing of university query to vectorstore

```
▶ (venv) ➔ AdaptiveRagChatbot git:(jivan) ✘ python main.py

Chroma db loaded from memory
Enter a question: Courses offered in harvard university

Route Question

SOURCE:vectorstore

RETRIEVE

Check Document relevance to question

Document is relevant

Document is not relevant to the question

Document is not relevant to the question

Document is relevant

Assess graded documents

Decision: Generate answer

GENERATE

Check hallucinations

Generation is grounded in the documents

Grade generation vs question

Decision: Generation answers the question
Courses offered in harvard university Harvard University offers various courses, including Law, Business & Management, Medicine, Engineering, Public Health, Education, Social Sciences, and Humanities. The university also provides graduate programs such as Master of Laws, Master of Business Administration, Doctor of Medicine, and PhD in Social Sciences and Biological Sciences. Specific undergraduate courses are not mentioned in the provided context.

(venv) ➔ AdaptiveRagChatbot git:(jivan) ✘
```

Routing of specific professor query

```
(venv) ➔ AdaptiveRagChatbot git:(jivan) ✘ python main.py
```

Chroma db loaded from memory

Enter a question: what is research areas of professor Aarti Singh of CMU

Route Question

SOURCE:professor_search

Route Professor Query

SOURCE:json_data

Professor Search from JSON

Professor Search by extracting website from json

Aarti Singh Carnegie Mellon University

WEBSITE : <http://www.cs.cmu.edu/~aarti>

SCRAPED DATA FROM WEBSITE : Error fetching website content: HTTPConnectionPool(host='www.cs.cmu.edu', port=80): Max retries exceeded with url: /~aarti (Caused by NewConnectionError(at 0x70acc8529950): Failed to establish a new connection: [Errno 111] Connection refused'))

<class 'str'>

name='Aarti Singh' title='Associate Professor' department='Computer Science' email='aarti@cs.cmu.edu' phone='412-268-3066' office='Gates Hall 7111' research_interests='Machine Learning' 'Aarti Singh is an associate professor in the Computer Science Department at Carnegie Mellon University.' publications='https://www.cs.cmu.edu/~aarti/publications.html' website='http://www.cs.cmu.edu/~aarti'

Route according to JSON Results

SOURCE:format_search_results

Format Search Results

what is research areas of professor Aarti Singh of CMU Machine Learning, Data Mining, Computational Biology.

```
(venv) ➔ AdaptiveRagChatbot git:(jivan) ✘
```

Routing of general professor query

```
(venv) ➔ AdaptiveRagChatbot git:(jivan) ✘ python main.py

Chroma db loaded from memory
Enter a question: Professor doing research in reinforcement learning and robotics

Route Question

SOURCE:professor_search

Route Professor Query

SOURCE:web_search

Professor Web Search

Format Search Results
Professor doing research in reinforcement learning and robotics The professors doing research in reinforcement learning and robotics are:

1. Animesh Garg - Assistant Professor at University of Toronto
2. Pieter Abbeel - Professor at UC Berkeley
3. Dr. Simchowitz - Incoming Assistant Professor, jointly appointed in MLD and CSD
4. Manuela Veloso - Emeritus, Herbert A. Simon University Professor, Machine Learning & Computer Science, School of Computer Science

These professors are working on various projects related to reinforcement learning and robotics, including mobile-manipulation, humanoid robotics, and generative AI.
```

Implementation Details

- Data Collection : Scraping using Scrapy/BeautifulSoup
- Data Preprocessing : Python, Numpy, Pandas
- RAG : Open source LLM (Using Groq) and Embeddings, Vector DB(Chroma) , Langchain and Langgraph
- Agentic Workflow : CrewAI / SmolaAgents, LLM (using provider like Groq)
- SOP Review : Using LLMs
- Frontend : HTML, CSS, Javascript
- Backend : FastAPI or Flask
- Database : Chroma (vectorstore), PostgreSQL(chat sessions) ,SQLAlchemy(ORM)

Remaining Tasks

- Refining the Langgraph workflow, and tools implementation
- Expanding and curating the dataset for more effective querying
- SOP Review Implementation
- Frontend and Backend Development
- Testing with various embedding and chunking strategies

Enhancements

- Automatic data collection and updation
- Voice based input for chat features
- Interactive map to show university locations.
- Plagiarism check for SOPs
- Multi Language support

THANK YOU