

Machine Learning Take Home Assignment

Objective

Develop a Retrieval Augmented Generation (RAG) system that can answer questions based on content from specified websites.

Task Deliverables

1. RAG System

- Implement a RAG system capable of answering user questions using information from provided websites.
 - Websites that will be focused on
 - <https://huyenchip.com/2024/07/25/genai-platform.html>
 - <https://lilianweng.github.io/posts/2024-07-07-hallucination/>
 - <https://jina.ai/news/what-is-colbert-and-late-interaction-and-why-they-matter-in-search/>
 - <https://quoraengineering.quora.com/Building-Embedding-Search-at-Quora>
- Ensure the system can accurately retrieve relevant information and generate coherent responses.
- Ensure that given any link the system is able to retrieve information from the **provided link** and can store it in candidate's choice of vector database for retrieval

2. API Development and Deployment

- Create an API Endpoint for the RAG system with basic setup.

- User can give website link and it will extract information from the website and store it in a vector database

- Endpoint = `api/v1/index`

- Sample Request for indexing:

```
{
  "url": ["https://example.com/article"]
}
```

- Sample Response for indexing:

```
{
  "status": "success",
  "indexed_url": ["https://example.com/article"]
  "failed_url": null
}
```

- User can give question and get answer from the system in Chat Format.

- Sample Request for chat format

```
{
  "messages": [
    {
      "content": "What are extrinsic hallucinations",
      "role": "user"
    },
    {
      "content": "Extrinsic hallucinations refer",
      "role": "assistant"
    },
    {
      "content": "How will it benefit me",
      "role": "user"
    }
  ]
}
```

```
}
```

- Endpoint `api/v1/chat`
- Sample Response for chat format

```
{  
  "response": [  
    "answer": {  
      "content": "...",  
      "role": "assistant"  
    },  
    "citation": ["link"]  
  ]  
}
```

- Deploy the API on a provider of your choice (e.g., AWS, Google Cloud, Heroku, railway app,).
- Provide clear documentation on how to use the API.

3. Citation Mechanism

- Implement a citation system that references the source of information in the generated answers.
- Ensure citations are accurate and link back to the original content on the websites.

4. Authentication (Bonus)

- Implement a basic authentication system for API access.
- Document the authentication process and provide example usage.

5. User Interface (Bonus)

- Develop a minimalistic UI using Streamlit, Gradio, or any framework of your choice.
- The UI should allow users to input questions and display the answers with citations.

Evaluation Criteria

- Accuracy and relevance of the generated answers
- Proper implementation and deployment of the API
- Correct citation of sources in the answers
- Code quality and documentation
- Bonus points for authentication and UI implementation

Submission Guidelines

- Provide a Link to Private GitHub repository with your code and documentation.
- Include a README file with setup instructions and API usage examples.
- Submit a brief report explaining your approach, challenges faced, and potential improvements.

Good luck with your assignment!