GenAI PS

By Overlayy

Institute: IIT Kanpur

Submitted by:

1. Pratibha Gupta
2. Dhiraj Pareek
3. Arpit Nigam
4. Prashik Ganer

Procedure

As part of our group's initiative to develop an automated system for generating relevant questions from website content, initially, we utilized BeautifulSoup to perform **website scraping**. This involves sending HTTP requests to retrieve the HTML content of a webpage, parsing it to extract all hyperlinks, and saving these URLs

Following this, we did **webpage content retrieval**. For each extracted URL, we will fetch the HTML content, extract the main text, and store it in a JSON file along with the URL. For the **question generation phase**, we employed the Gemini API (text-embedding-004 model) to produce ten concise questions for each webpage based on its content. We began by retrieving the previously saved textual responses for each webpage. Utilizing the Gemini API, we generated questions that were both relevant to the content and limited to fewer than 80 characters each. This also ensured that the questions were precise and meaningful. To facilitate easy access and organization, we stored the generated questions in structured JSON files, with each file corresponding to a specific URL. Each JSON(questions.json) file includes the URL and its associated set of ten questions, allowing for efficient retrieval and use in subsequent stages of our project.