Cognitive Summarization and Question Generation System – Workflow

1. Data Collection

The system begins with collecting educational material in the form of PDF documents, lecture notes, textbooks, or scanned study materials. These documents are uploaded by students or educators through the platform’s upload interface.

2. Data Preprocessing

Uploaded documents undergo OCR (if scanned), followed by text extraction. The extracted content is cleaned by removing stop words, special characters, and unnecessary whitespace. Tokenization, sentence segmentation, and lemmatization are also applied.

3. User Input Interface

Users can either select the entire document for processing or highlight specific sections they want summarized or question-generated. Input preferences like summary length or question difficulty can also be configured.

4. NLP Techniques Applied

The backend utilizes transformer-based models like BART or T5 for extractive and abstractive summarization. For question generation, a sequence-to-sequence model fine-tuned on educational QA datasets is used. Named Entity Recognition (NER) and keyphrase extraction aid in generating context-aware questions.

5. Output Generation

The final output includes: - A summarized version of the uploaded content. - A set of automatically generated questions (MCQs, True/False, Short Answer). Outputs are downloadable and optionally saved to user profiles for future study.

6. End Users

The primary users of this system include: - Students preparing for exams who need quick content revision and self-assessment. - Teachers and educators who want to create quick summaries or test questions. - Educational platforms aiming to enhance interactivity and learning using NLP.