Topic: Education – specific: Classroom Engagement Assistant

Objective: To build an expert system for answering questions from text extracted from pdf

**Purpose: Enhances student engagement during lessons.**

Capabilities: Facilitates real-time Q&A, polls, quizzes, and class discussions, making lessons more interactive. It can also help teachers gauge student understanding by summarizing discussions or highlighting areas where students struggle.

Data collection (Thursday 2nd Oct - deadline)

* Extract pdf documents from google scholar – subjects: Computer science, Public Health, Geography
* Split the pdfs into images
* Annotate the images for training and validation
* Yolo/Detectron –

Steps for data preparation

* Text extraction
* Text labeling
* Data cleaning
* Training and Model fine-tuning
* Answer questions

**Text and Data Extraction**

* Use **PDF parsing tools** to extract the required information from PDFs:
  + **Raw text** for training your assistant on topics, concepts, and definitions.
  + **Image captions** to ensure the assistant can reference or explain images.
  + **Table data** to handle structured information like statistics or datasets.
  + **Form data** (if the PDFs have fillable forms) to simulate interactive activities or assessments

**Data Structuring**

* **Split the extracted content** into logical sections:
  + Use **headings** or **keywords** to break content into manageable topics (e.g., programming concepts, epidemiology, geographical climate zones).
  + Ensure **image captions** are linked to the associated images and stored properly.
  + Organize **table data** into structured formats (e.g., rows and columns) for easy integration into quizzes or exercises.

**Data Labeling**

* **Label the extracted content** by topic or section:
  + Label sections based on their subject matter (e.g., "Data Structures", "Disease Prevention", "Geological Formations").
  + For **tables**, ensure the data is categorized (e.g., "Population Statistics", "Algorithm Analysis").

**Dataset Creation**

* **Organize the extracted and labeled data** into a structured dataset:
  + Create a **Q&A format** from the content to train your model. For instance, you can generate questions from text, tables, or image captions.
  + Store the dataset in a format suitable for training machine learning model.

**Refinement**

* + Based on the assistant's performance, continue to refine the model by adding more examples or improving the dataset.
  + Ensure that the assistant provides clear and accurate responses for each of the subjects and data types (text, captions, tables, forms).
  + (e.g., **CSV**, **JSON**).

Data Areas

**Computer Science**

1. **Algorithms and Data Structures**
   * Sorting, searching, trees, graphs, hashing, dynamic programming.
2. **Operating Systems**
   * Process management, memory management, file systems, concurrency.
3. **Computer Networks**
   * TCP/IP, network protocols, security, routing, network topologies.
4. **Artificial Intelligence and Machine Learning**
   * Neural networks, decision trees, reinforcement learning, NLP.
5. **Database Management Systems (DBMS)**
   * SQL, NoSQL, database design, normalization, transactions.
6. **Software Engineering**
   * Software development life cycle (SDLC), agile methodologies, version control.
7. **Programming Paradigms**
   * Object-oriented programming, functional programming, imperative and declarative paradigms.
8. **Cybersecurity**
   * Cryptography, firewalls, intrusion detection systems, malware analysis.
9. **Cloud Computing**
   * Virtualization, SaaS, IaaS, PaaS, cloud security, DevOps.
10. **Human-Computer Interaction (HCI)**
    * User interface design, UX principles, interaction techniques, usability.

**Public Health**

1. **Epidemiology**
   * Disease transmission, incidence/prevalence, surveillance systems, outbreak investigation.
2. **Global Health**
   * Communicable diseases (HIV, tuberculosis, malaria), global health initiatives, WHO programs.
3. **Health Promotion and Disease Prevention**
   * Behavioral health interventions, vaccination programs, community health strategies.
4. **Health Policy and Management**
   * Healthcare systems, health economics, healthcare reform, policy analysis.
5. **Environmental Health**
   * Pollution, climate change, sanitation, occupational health hazards.
6. **Biostatistics**
   * Statistical methods in health research, hypothesis testing, regression analysis, survival analysis.
7. **Social Determinants of Health**
   * Income inequality, education, housing, social justice and equity in healthcare.
8. **Nutrition and Public Health**
   * Malnutrition, food security, obesity, dietary interventions.
9. **Maternal and Child Health**
   * Prenatal care, infant mortality, child vaccination, reproductive health.
10. **Infectious Disease Control**
    * Quarantine, contact tracing, antibiotic resistance, vaccination and herd immunity.

**Geography**

1. **Physical Geography**
   * Climate systems, geomorphology, biogeography, hydrology, landforms.
2. **Human Geography**
   * Population geography, urbanization, migration, cultural geography, geopolitics.
3. **Geographical Information Systems (GIS)**
   * Spatial analysis, remote sensing, cartography, map interpretation.
4. **Climate Change**
   * Global warming, greenhouse gases, climate models, impact on ecosystems and human populations.
5. **Sustainable Development**
   * Resource management, renewable energy, environmental conservation, sustainability initiatives.
6. **Economic Geography**
   * Global trade, economic globalization, industrial location, agriculture, economic disparities.
7. **Urban Geography**
   * Urban planning, megacities, smart cities, rural-urban migration, urban infrastructure.
8. **Environmental Hazards and Risk Management**
   * Natural disasters (earthquakes, hurricanes, floods), mitigation strategies, disaster risk management.
9. **Political Geography**
   * Territorial boundaries, geopolitical conflicts, international relations, political divisions.
10. **Cultural Geography**
    * Languages, religions, cultural landscapes, identity, and spatial patterns of cultural practices.