Project Synopsis

Title: Design and Development of a Biomedical Question-Answering System using Domain-Specific Transformers

1. Problem Statement

The biomedical domain generates an enormous volume of research articles and clinical content daily. Medical practitioners, researchers, and students often struggle to find precise answers to their queries in this unstructured literature. Traditional search engines retrieve documents, but do not answer specific biomedical questions. This project aims to build a **Biomedical Question-Answering (QA)** system that can return **concise, evidence-backed answers** from biomedical corpora, thereby supporting **literature review, clinical decision-making**, and **academic learning**.

2. Objective

To build an end-to-end pipeline that can understand natural language health-related queries and return contextually relevant answers from a biomedical corpus using **transformer-based models (BioBERT/PubMedBERT)** fine-tuned for biomedical QA tasks.

3. Scope of Work

- Build a **retrieval-based** QA system using transformer models
- Use either the **BioASQ QA dataset** or **PubMed abstracts** as the knowledge source
- Fine-tune a pre-trained QA model like **BioBERT** or **PubMedBERT**
- Design a simple web or command-line interface for testing queries
- Evaluate using QA metrics: **Exact Match (EM)** and **F1-score**

4. Dataset Description

Primary Dataset: BioASQ

- A benchmark biomedical QA dataset provided by the BioASQ challenge
- Contains factoid and list-type questions with gold-standard answers from PubMed abstracts

- Downloadable in JSON format: http://bioasq.org

5. Tools & Technologies

- **Hugging Face Transformers (BioBERT, PubMedBERT)**
- **Python (Transformers, Datasets, NLTK, scikit-learn)**
- **BioASQ JSON format handling**
- Optional frontend: **Streamlit / Gradio / Flask**

6. Course Outcome Mapping

- **CO2:** Implementation of biomedical NLP algorithms using transformer models
- **CO3:** Designing a domain-specific QA system for the healthcare/pharma sector

7. Expected Outcome

- A working biomedical QA system capable of:
    - Accepting natural language queries (e.g., "What causes Alzheimer's disease?")
    - Retrieving the relevant biomedical context
    - Extracting a concise, evidence-based answer
- A comparative evaluation of different transformer models on biomedical QA
- Insight into the challenges of domain-specific QA (e.g., terminology, ambiguous phrasing)

8. Conclusion

This project bridges the gap between complex biomedical text and accessible healthcare knowledge by enabling automated question answering. It demonstrates practical application of NLP in healthcare and aligns with current research in **clinical informatics** and **digital health tools**. The project contributes toward improving information accessibility for researchers, students, and even patients in the future.