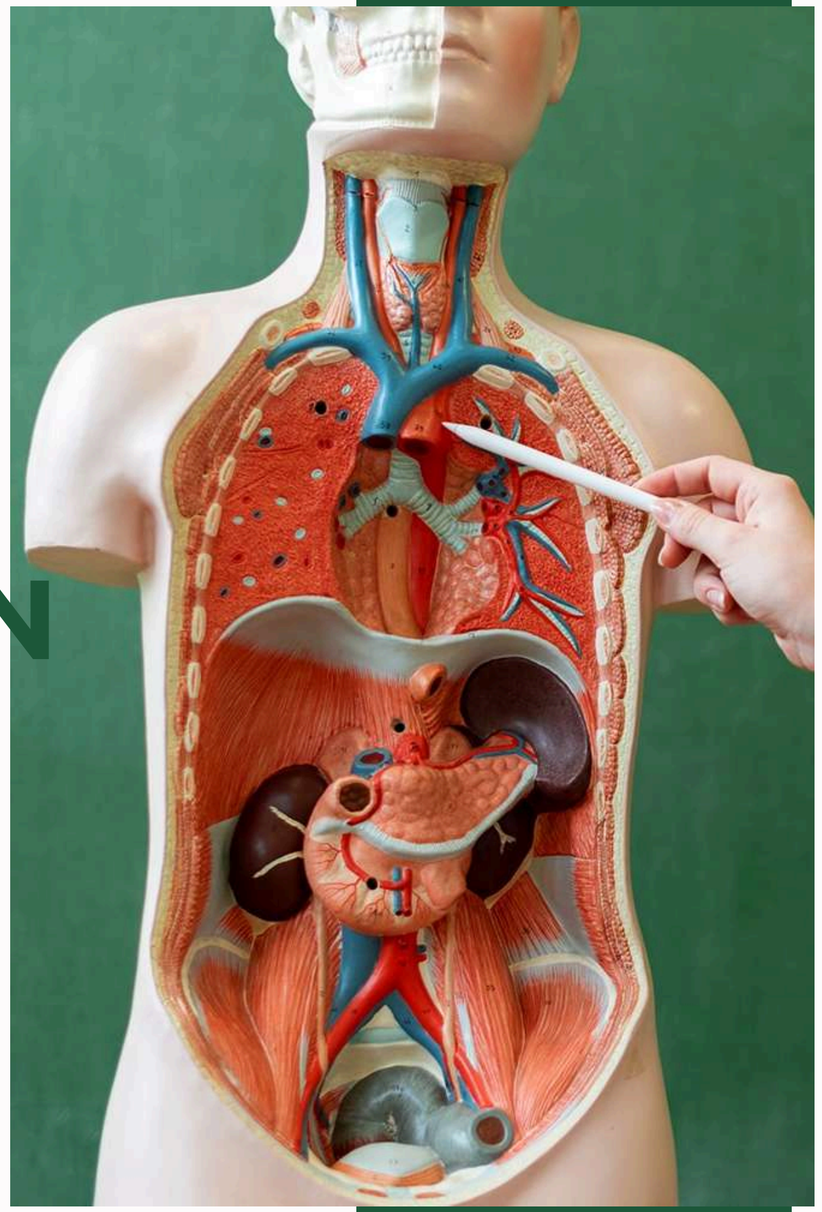
BIOMENTOR - PERSONALIZED E-LEARNING PLATFORM FOR A/L BIOLOGY SUBJECT STUDENTS IN SRI LANKA

24-25J-257



SLIIT
FACULTY OF COMPUTING

Our Team

Srirajan G. A
IT21375132

Dharane.S
IT21068478

Sujitha.S
IT21264634

Sajeevan.S
IT21204302

N.Thayaparan
External Supervisor

SLIIT
FACULTY OF COMPUTING

Introduction

01 Background

02 Research Problem

03 Research Objectives

04 Overall System Diagram

SLIIT
FACULTY OF COMPUTING

Background

This project aims to create an engaging and effective learning environment that caters to individual learning styles and promotes continuous improvement through detailed feedback and performance tracking.

We’re focusing on this project to provide personalized learning in A-Level Biology using approved government resources. Our advanced technologies aim to offer tailored, engaging experiences that enhance retention and readiness.

SLIIT
FACULTY OF COMPUTING

Research Problem

01

A/L biology students in Sri Lanka struggle with memorizing complex biological terms and their pronunciation. There is a need for an interactive tool that combines spaced repetition with detailed feedback to enhance vocabulary retention and accuracy.

02

Students struggle with extensive biology texts, and existing tools don't effectively extract key concepts or support auditory learners. There is also a lack of tools that provide targeted summaries for exam preparation. A solution is needed to offer concise, accurate content that supports diverse learning styles and aligns with educational standards.

03

Static MCQ platforms do not adapt to students' abilities, leading to ineffective practice and poor identification of knowledge gaps. The objective is to develop an adaptive quiz system that adjusts question difficulty, offers detailed performance analytics, and provides targeted feedback.

04

Existing evaluation systems for biology responses often lack comprehensive feedback and actionable recommendations. The problem is to design a platform that provides accurate answers, detailed feedback, and additional study resources to support student improvement.

Objectives

Objective 01

Biology vocabulary memorization tool using digital flashcards and spaced repetition, providing personalized feedback and adaptive learning paths.

Objective 02

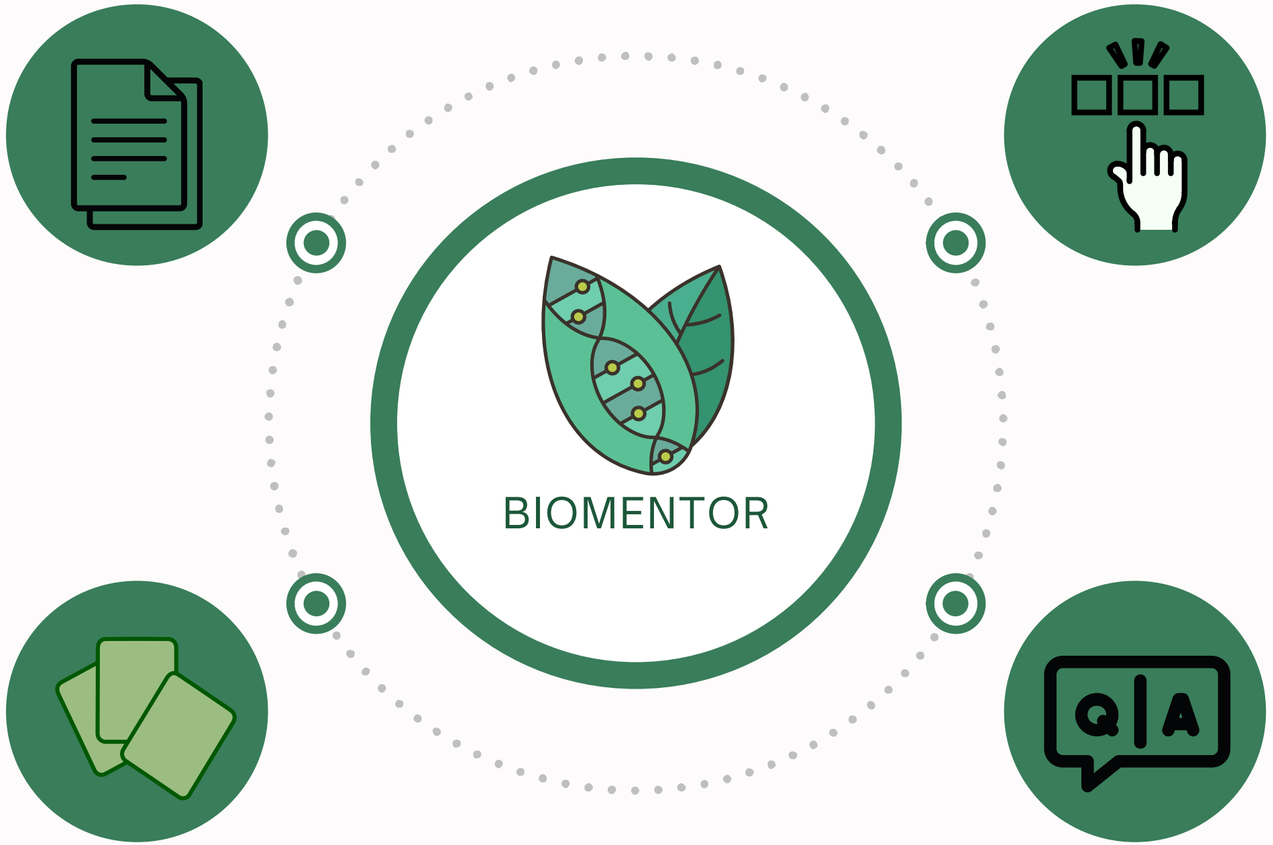
Summarization tool that generates concise, topic-based summaries from uploaded documents and searches through resources to produce summaries on specific topics, with customizable word counts and voice output for diverse learning styles.

Objective 03

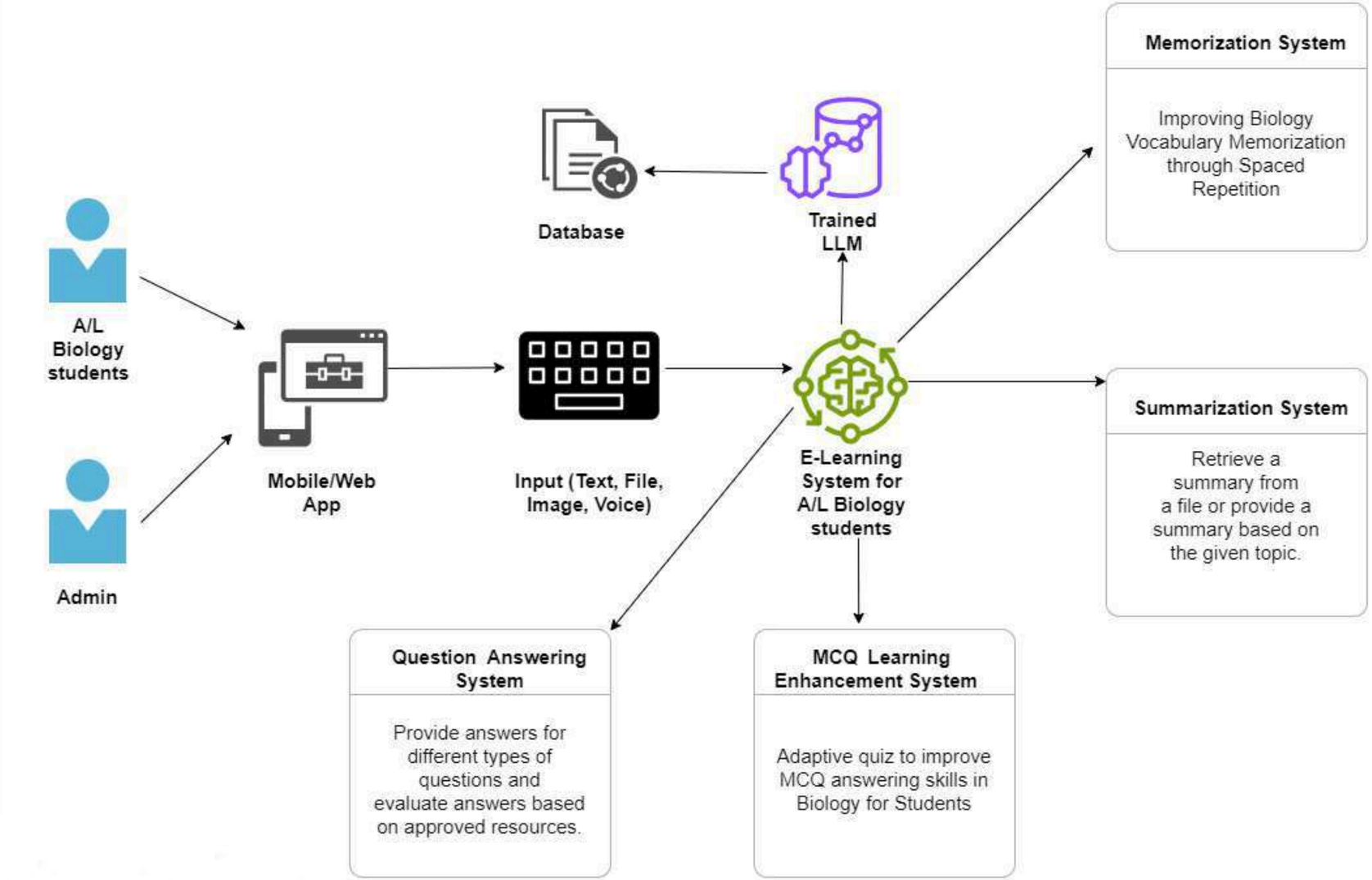
An adaptive quiz platform that dynamically adjusts question difficulty based on student performance, offering targeted practice and detailed performance analysis to enhance learning outcomes.

Objective 04

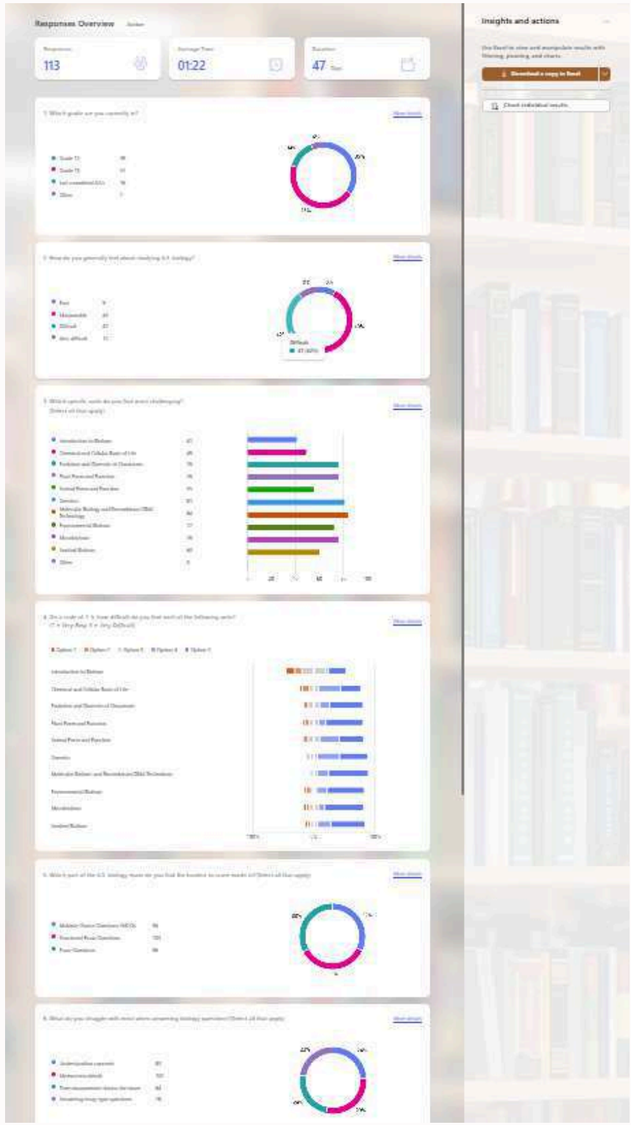
Q & A platform to generate answers for provided questions and evaluate responses, offering feedback and personalized study resources to improve student understanding and performance.



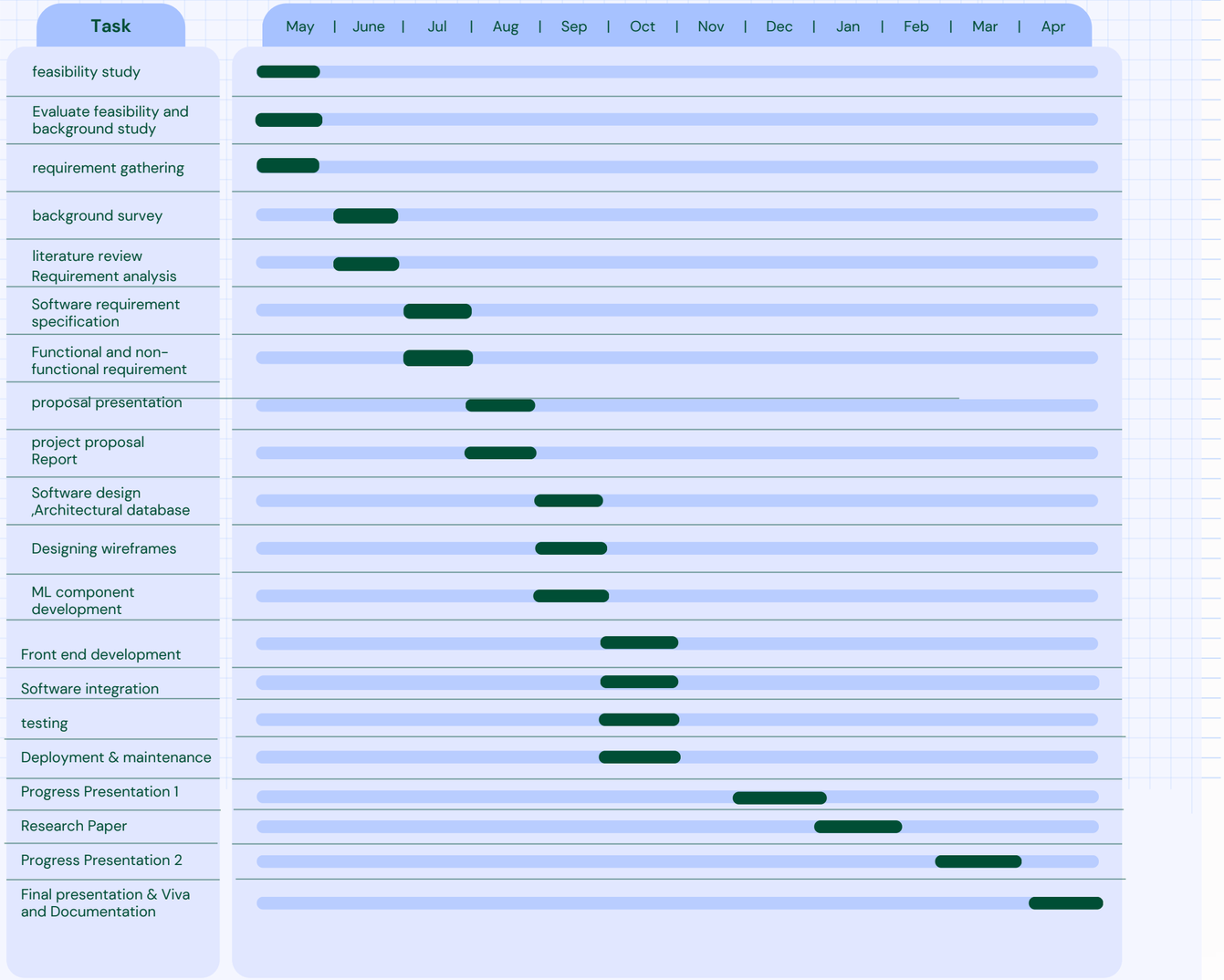
Overall System Diagram



Survey



Gantt Chart


Task
May
June
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
feasibility study
Evaluate feasibility and background study
requirement gathering
background survey
literature review Requirement analysis
Software requirement specification
Functional and non-functional requirement
proposal presentation
project proposal Report
Software design ,Architectural database
Designing wireframes
ML component development
Front end development
Software integration
testing
Deployment & maintenance
Progress Presentation 1
Research Paper
Progress Presentation 2
Final presentation & Viva and Documentation

BIOMENTOR

Cost Management Plan

Typs	Cost
Internet use and web hosting	LKR.10,000.00
Training Cost	LKR.30,000.00
Publication Cost	LKR.70,000.00
Stationery	LKR.1,000.00
Total	LKR.111,000.00

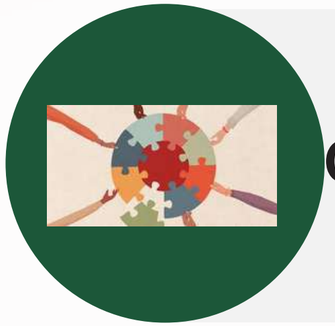
SLIIT FACULTY OF COMPUTING

BIOMENTOR

Commercialization

Advertising

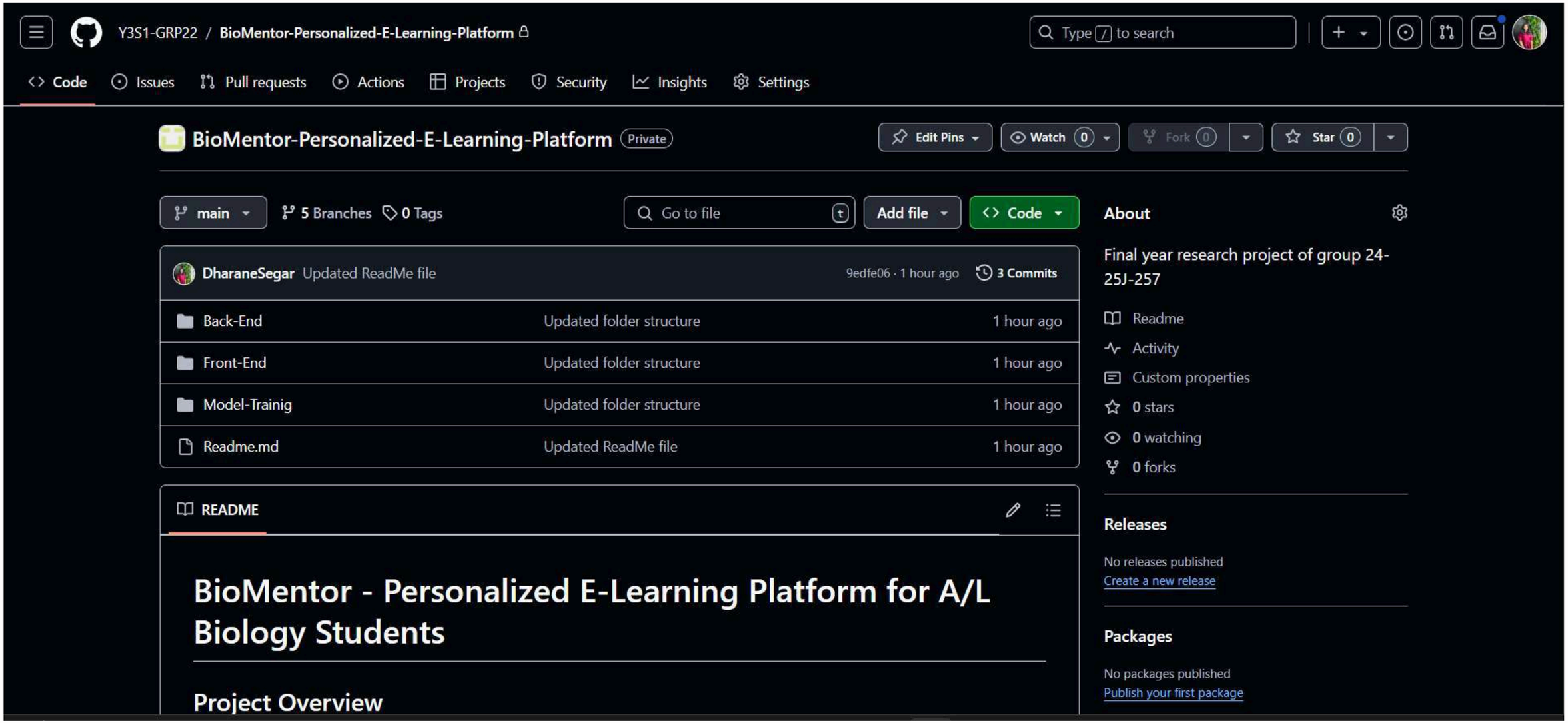
Subscription



Collaboration

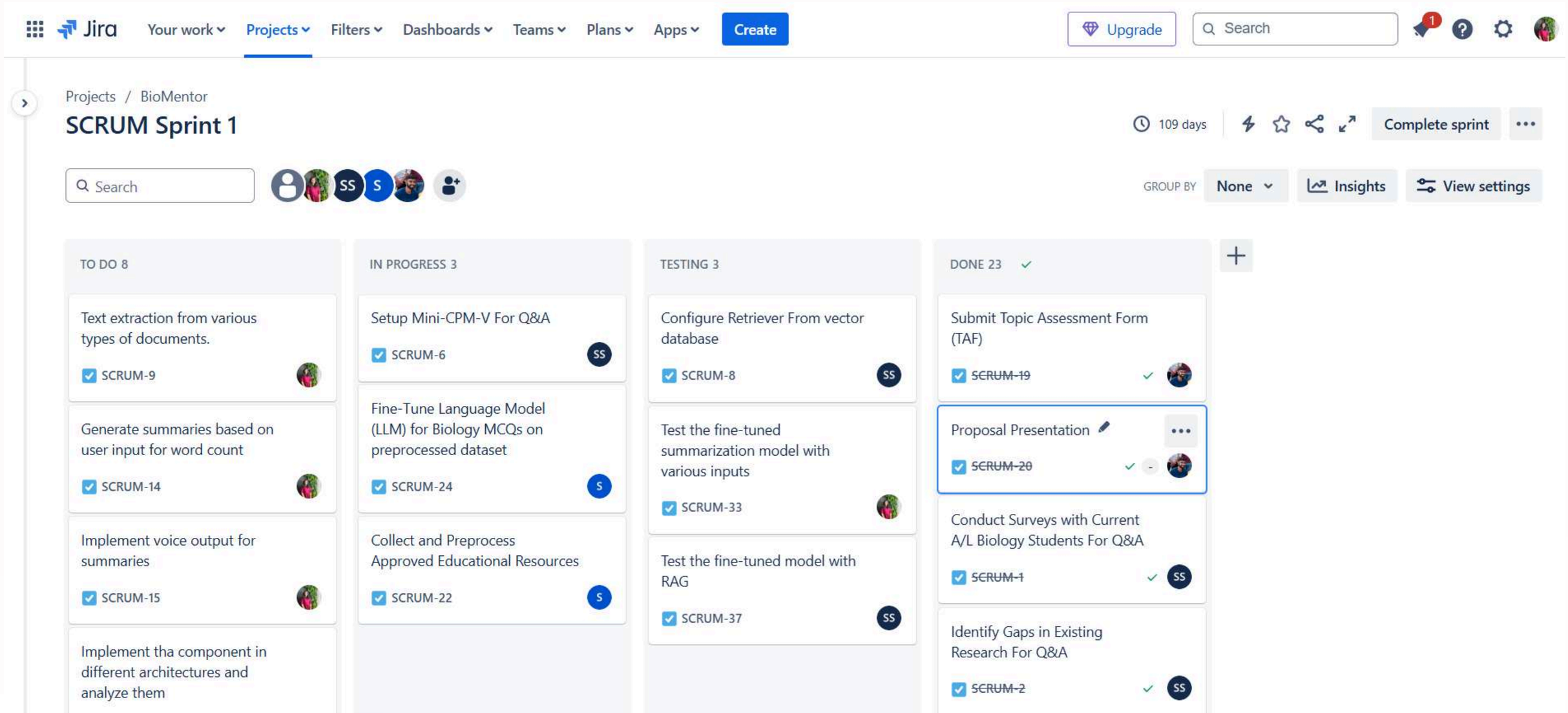
SLIIT
FACULTY OF COMPUTING

GitHub Details



SLIIT
FACULTY OF COMPUTING

Project Management



Risk Mitigation

Risk	Trigger	Owner	Response	Resource Required
Risk with respect to the Project Team				
Minor delays in deliverables	Small miscommunications or changes in priority	Project Leader	Discuss with team to identify any small delays. Adjust timelines as needed.	Updated Project Schedule Plan/Gantt Chart
Risk with respect to the Panel/Supervisor(s)				
Slight misalignment on expectations	Minor changes in requirements	Project Leader	Confirm and clarify expectations through quick discussions. Update project documents if necessary.	Meeting Notes Updated Project Plan
Panel preferences shift slightly	Panel's evolving preferences or needs	Project Leader	Seek feedback regularly to stay aligned with expectations. Make small adjustments as needed.	Meeting Logs Updated Deliverables
Risk with respect to Project Execution				
Minor technical hiccups	Occasional bugs or troubleshooting	Project Leader	Allocate time for quick debugging and testing. Work with the team to address issues.	Debugging Tools Technical Documentation

IT21375132

Srirajan G.A
Software Engineering

IMPROVE BIOLOGY VOCABULARY MEMORIZATION THROUGH SPACED REPETITION

Introduction

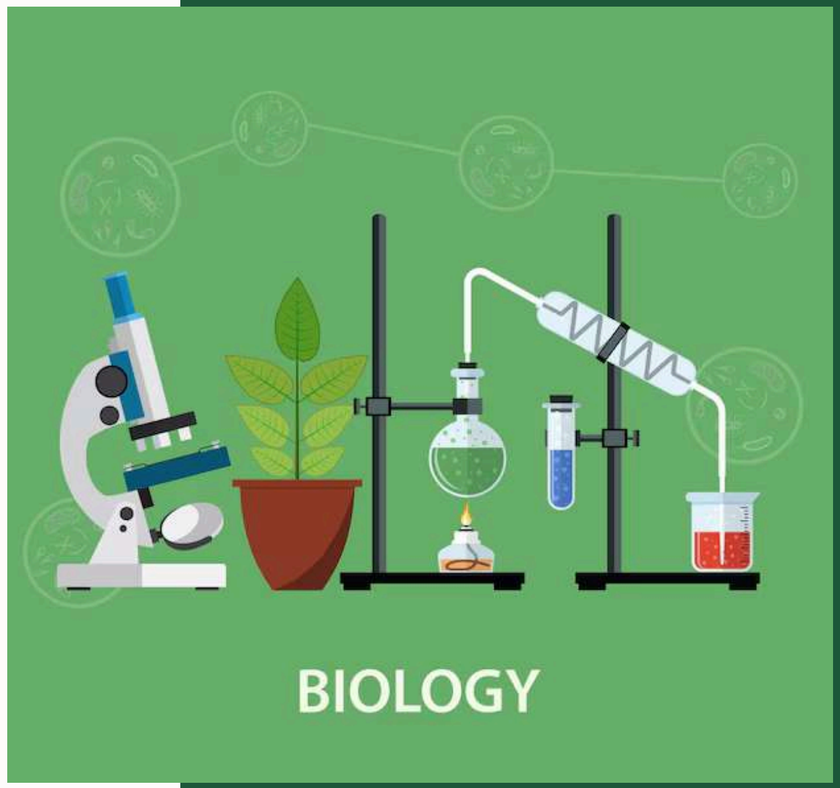
01 Background

02 Research Question

03 Research Gap

04 Main and Sub Objectives

05 Methodology



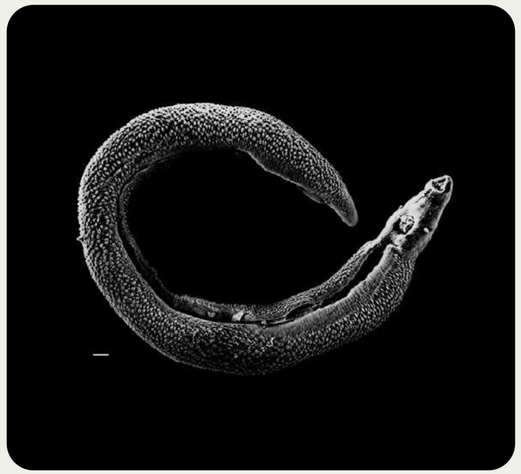
SLIIT FACULTY OF COMPUTING

BACKGROUND

Students struggle with retaining large amounts of biology vocabulary due to the natural forgetting curve, where newly learned information is quickly forgotten if not reviewed.

An innovative system is needed to leverage cognitive science principles, such as spaced repetition, to counteract the forgetting curve.

BACKGROUND



Schistosomiasis

Schistosomiasis is an acute and chronic disease caused by parasitic worms

RESEARCH PROBLEM

01

How can we customize the spaced intervals based on the performance of the user and the difficulty of the vocabulary to maximize memory retention among Advanced Level Biology students?

How can we incorporate multi-sensory spaced repetition techniques on the memorization of biology vocabulary to maximize the memory retention among Advanced Level Biology students?

02

SLIIT FACULTY OF COMPUTING

IT21375132 |Srirajan G.A | 24-25J-257

Research Gap

	Personalized Spaced Repetition	*Customized for A/L Biology Students*	*Flash Cards*	*Multi-Sensory Techniques*
	✗	✗	✓	✓
duolingo	✗	✗	✗	✓
BIOMENTOR	✓	✓	✓	✓

OBJECTIVES

Objective 1

Create a spaced repetition model that adapts to individual user performance and adjusts review intervals based on the difficulty level of the vocabulary.

Objective 2

Incorporating multi-sensory elements to cater to different learning styles and enhance the memorization process.

Main Objective

To create an interactive application with flashcards that enhances biology vocabulary memorization through a custom spaced repetition model, which analyzes user performance and the difficulty of the questions and will repeat accordingly.

Objective 3

Incorporate gamification elements to maintain user motivation and engagement throughout the learning process

Objective 4

Implement a model to check the accuracy of the word that was entered.

IT21375132 |Srirajan G.A | 24-25J-257

Methodology

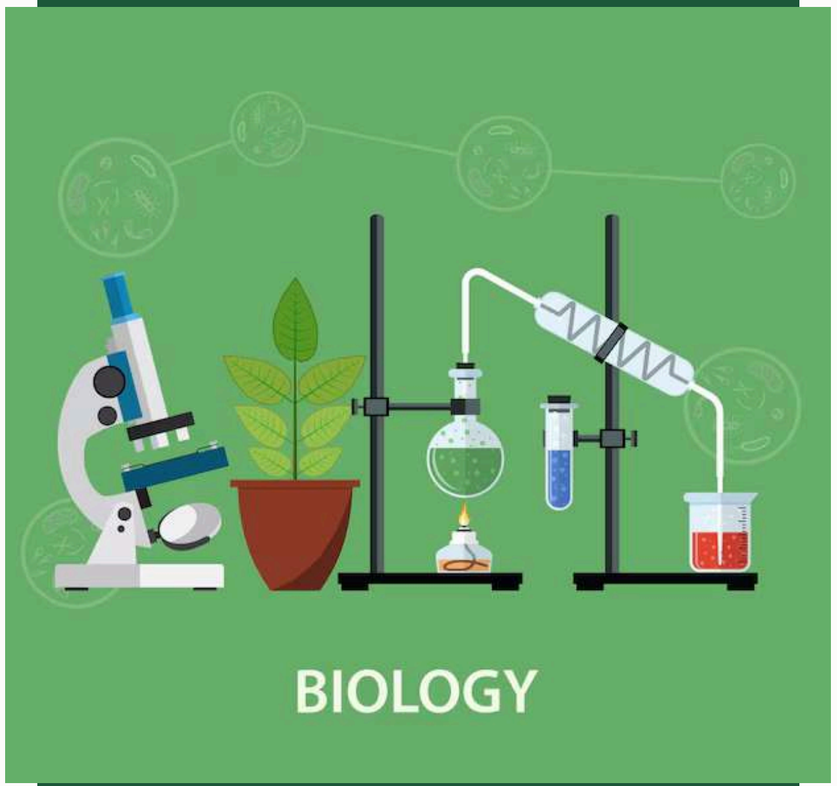
01 System Diagram

02 Tools and Technologies

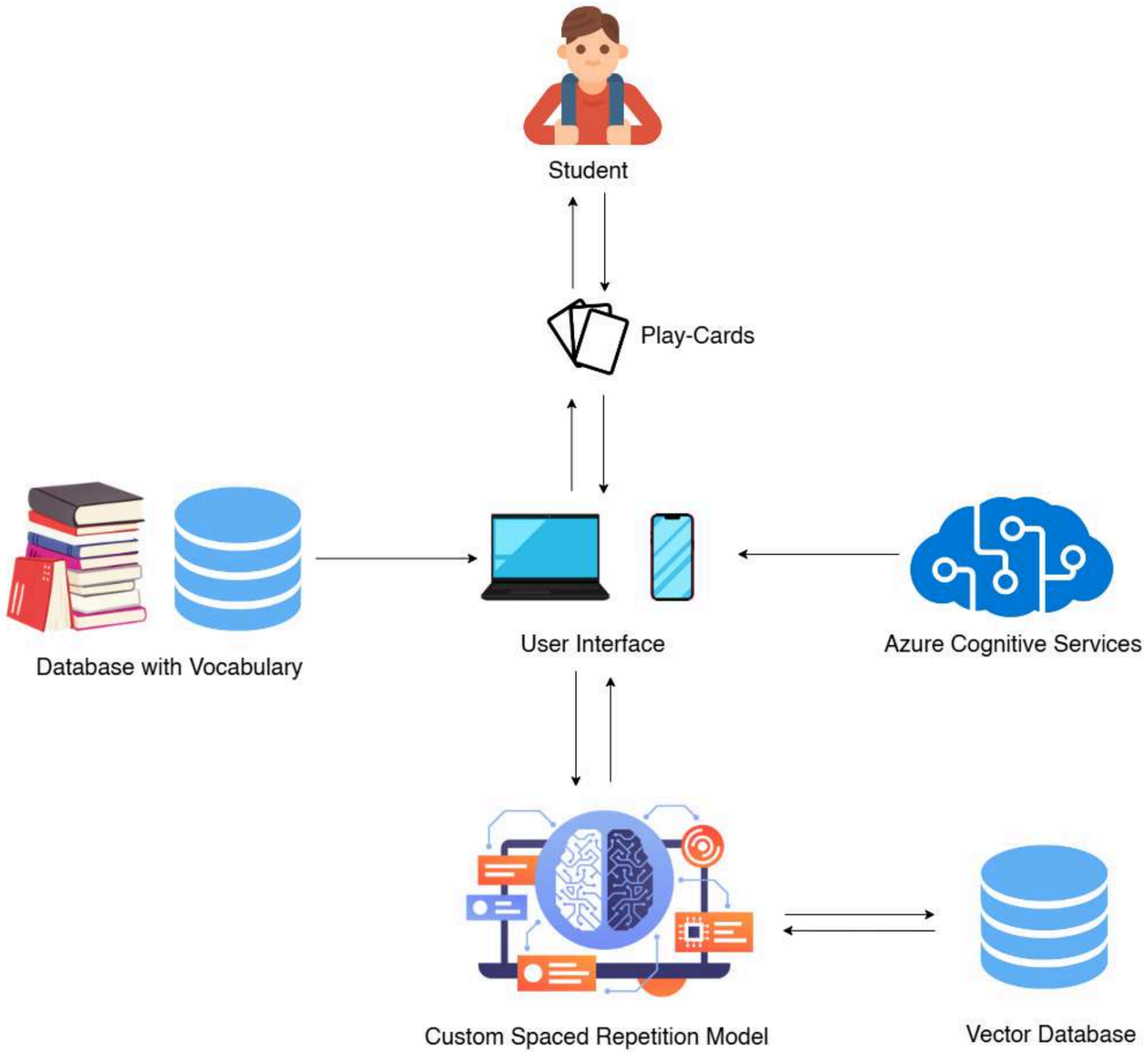
03 Requirements

04 Work Breakdown Structure

05 Gantt Chart



System Diagram



Tools & Technologies

Programming Languages

Python
Javascript

Database

Faiss
Mongo DB

Project Management

Jira

Frameworks

Flask
React
Tensorflow

Other tools

Git
Figma
Postman
Azure Cognitive Services

Requirements

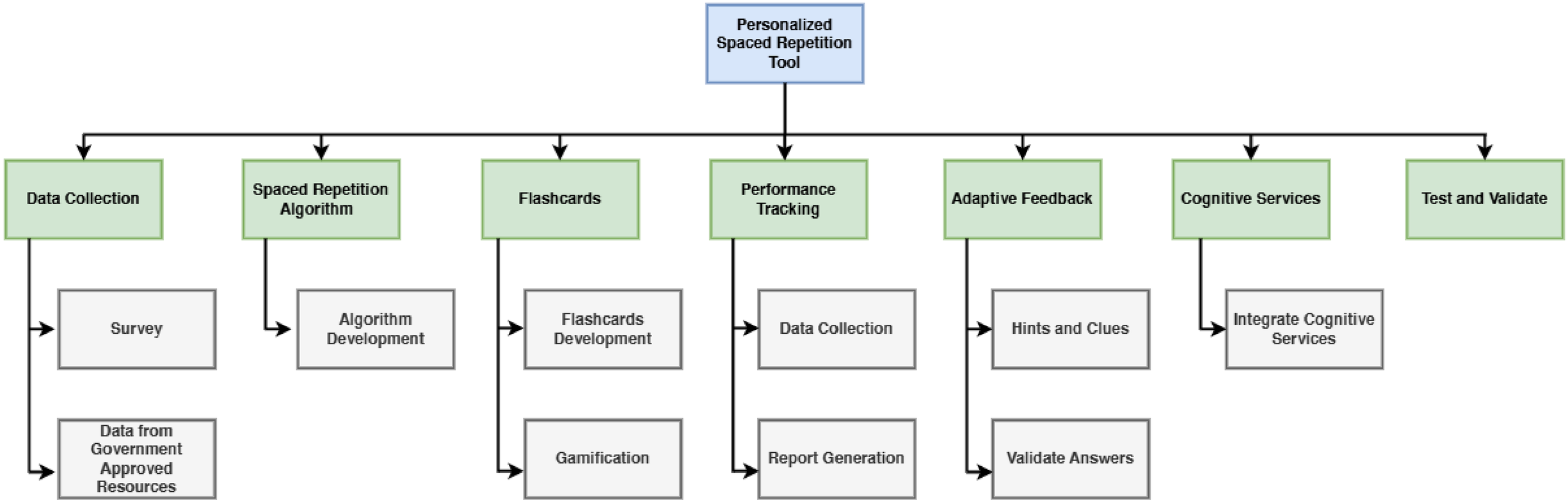
Functional

1. The tool should allow users to customize flashcard sets with multimedia elements, including images, audio, and text.
2. The spaced repetition model should adjust review intervals based on user performance and vocabulary difficulty.
3. The tool should track user performance, generate detailed reports, and provide instant, adaptive feedback.

Non-Functional

1. Compatibility
2. Accuracy
3. Performance
4. Usability
5. Availability

Work Breakdown Structure

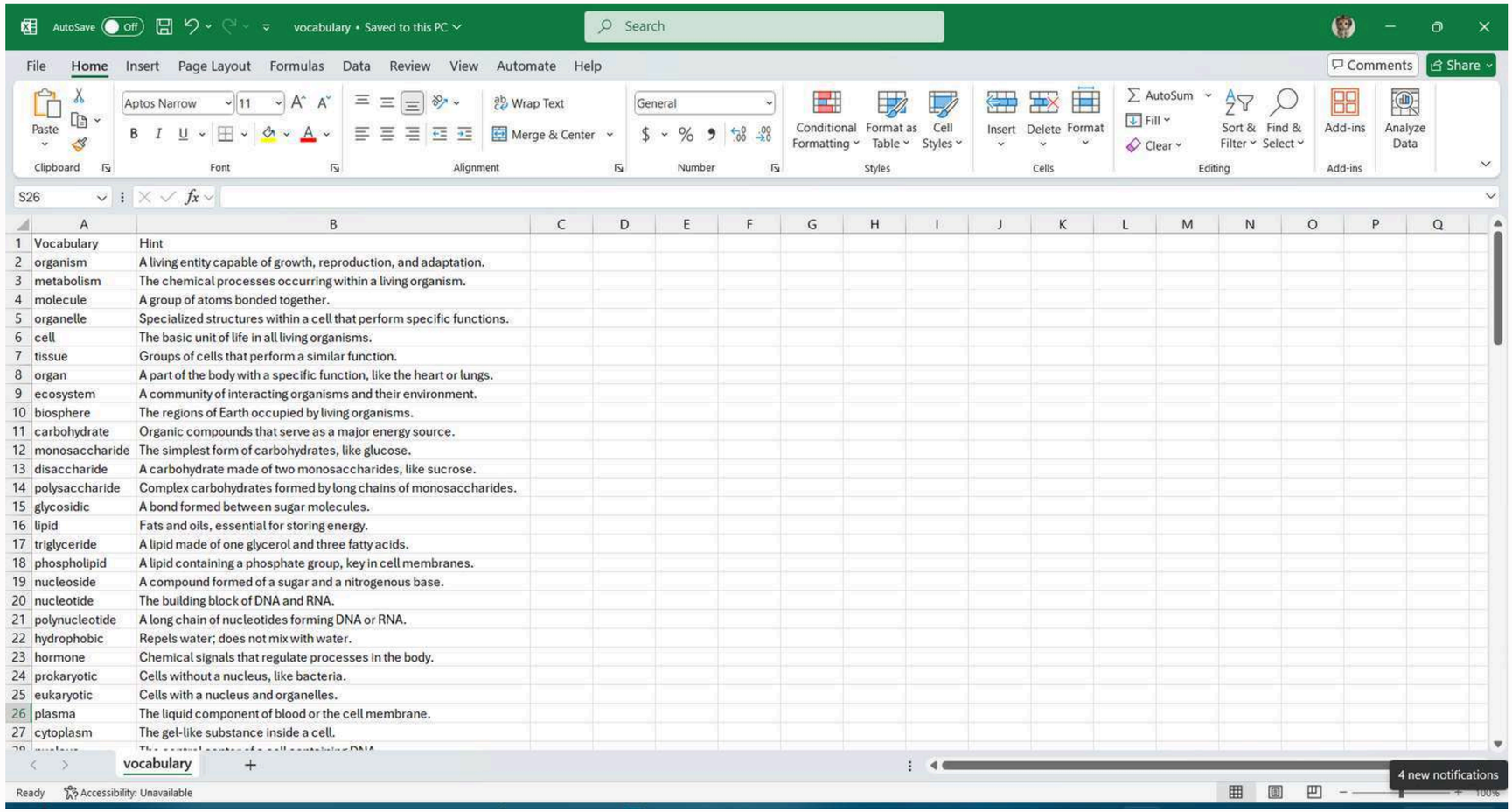


Completion of the project

SLIIT
FACULTY OF COMPUTING

IT21375132 |Srirajan G.A | 24-25J-257

Data Collection



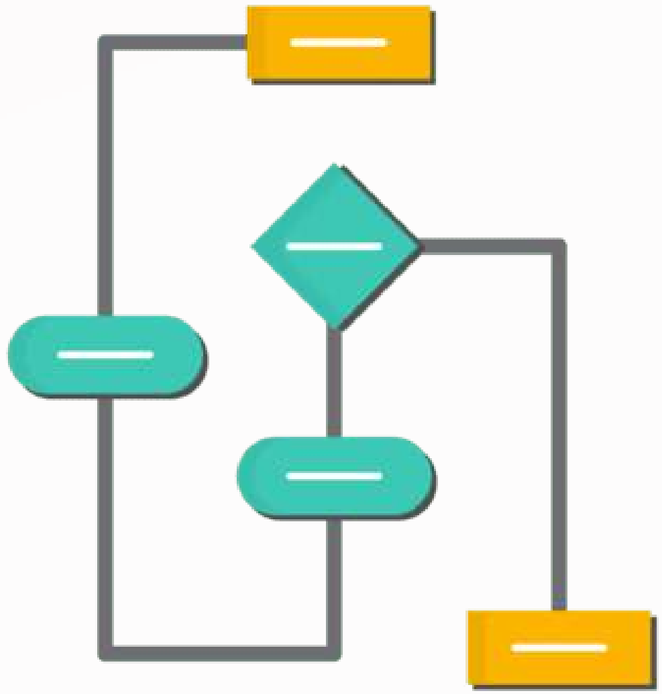
	A	B
1	Vocabulary	Hint
2	organism	A living entity capable of growth, reproduction, and adaptation.
3	metabolism	The chemical processes occurring within a living organism.
4	molecule	A group of atoms bonded together.
5	organelle	Specialized structures within a cell that perform specific functions.
6	cell	The basic unit of life in all living organisms.
7	tissue	Groups of cells that perform a similar function.
8	organ	A part of the body with a specific function, like the heart or lungs.
9	ecosystem	A community of interacting organisms and their environment.
10	biosphere	The regions of Earth occupied by living organisms.
11	carbohydrate	Organic compounds that serve as a major energy source.
12	monosaccharide	The simplest form of carbohydrates, like glucose.
13	disaccharide	A carbohydrate made of two monosaccharides, like sucrose.
14	polysaccharide	Complex carbohydrates formed by long chains of monosaccharides.
15	glycosidic	A bond formed between sugar molecules.
16	lipid	Fats and oils, essential for storing energy.
17	triglyceride	A lipid made of one glycerol and three fatty acids.
18	phospholipid	A lipid containing a phosphate group, key in cell membranes.
19	nucleoside	A compound formed of a sugar and a nitrogenous base.
20	nucleotide	The building block of DNA and RNA.
21	polynucleotide	A long chain of nucleotides forming DNA or RNA.
22	hydrophobic	Repels water; does not mix with water.
23	hormone	Chemical signals that regulate processes in the body.
24	prokaryotic	Cells without a nucleus, like bacteria.
25	eukaryotic	Cells with a nucleus and organelles.
26	plasma	The liquid component of blood or the cell membrane.
27	cytoplasm	The gel-like substance inside a cell.

https://www.nie.lk/pdffiles/tg/eALSyl%20Biology.pdf
https://www.nie.lk/pdffiles/other/eGr12OM%20BioResoBook.pdf

Spaced Repetition Algorithm Selection

Why SM-2?

- Scientifically proven retention.
- Highly customizable algorithm.
- Dynamically adjusts review schedules according to user performance.



SM-2 Algorithm

$$E' = E + 0.1 - (5 - Q) * (0.08 + (5 - Q) * 0.02)$$

$$\text{Interval}' = \begin{cases} 1 & \text{if repetitions} = 1 \\ 6 & \text{if repetitions} = 2 \\ \text{Interval} * E & \text{if repetitions} > 2 \end{cases}$$

If Q < 3

1. Reset repetitions to 0
2. Set interval to 1

SM-2 Algorithm

```python
import datetime
import math
import random
import itertools
import csv

# Define the time format for consistent display
time_fmt = "%Y-%m-%d"

class Card:
    """
    Represents a single flashcard used in spaced repetition.
    Attributes:
        - top: The front side of the card (e.g., question, term).
        - bot: The back side of the card (e.g., answer, definition).
        - time: The last review time.
        - repetitions: Number of successful repetitions.
        - interval: Days until the next review.
        - easiness: Easiness factor (affects interval growth).
    """
    def __init__(self, top, bot, time, repetitions=0, interval=1, easiness=2.5):
        self.top = top
        self.bot = bot
        self.time = time.replace(second=0, microsecond=0)
        self.repetitions = repetitions
        self.interval = interval
        self.easiness = easiness

    @property
    def is_new(self):
        """Determine if the card is new (has not been reviewed yet)."""
        return self.repetitions == 0

    @property
    def next_time(self):
        """Calculate the next review date based on the current interval."""
        return self.time + datetime.timedelta(days=math.ceil(self.interval))
```

```python
def repeat(self, quality, time):
    """
    Update the card's state using the SM-2 algorithm.
    Parameters:
        - quality: Review quality (0-5) based on user performance.
        - time: The current datetime.
    """
    assert 0 <= quality <= 5, "Quality should be between 0 and 5."

    # Adjust easiness factor based on quality
    self.easiness = max(1.3, self.easiness + 0.1 - (5 - quality) * (0.08 + (5 - quality) * 0.02))

    if quality < 3:
        # If performance is poor, reset repetitions and interval
        self.repetitions = 0
        self.interval = 1
    else:
        # If performance is acceptable, increment repetitions
        self.repetitions += 1
        if self.repetitions == 1:
            self.interval = 1  # First repetition interval
        elif self.repetitions == 2:
            self.interval = 6  # Second repetition interval
        else:
            # Subsequent intervals grow based on easiness factor
            self.interval *= self.easiness

    # Update the review time
    self.time = time

def __repr__(self):
    """Provide a human-readable representation of the card."""
    return (f"Card: {self.bot}, Next Review: {self.next_time.strftime(time_fmt)}, "
            f"Repetitions: {self.repetitions}, Interval: {self.interval:.2f}, Easiness: {self.easiness:.2f}")
```

Determine Review Quality

Why Levenshtein Distance Algorithm?

- Captures detailed differences between text inputs providing Accuracy.
- Error Handling allows scores and identifies common user mistakes.
- Customizable thereby can be used for Quantitative Scoring.

H	O		N	D	A	
H	Y	U	N	D	A	I

H	Y	U	N	D	A	I
H		O	N	D	A	

Levenshtein Distance Algorithm

```python
from Levenshtein import distance

def word_accuracy(original_word, entered_word):
    """
    Calculate the accuracy score between two words.

    Parameters:
    - original_word (str): The correct word.
    - entered_word (str): The word entered by the user.

    Returns:
    - int: Accuracy score between 0 (worst) and 5 (best).
    """

    # Normalize the words by converting to lowercase
    original_word = original_word.lower()
    entered_word = entered_word.lower()

    # Calculate Levenshtein distance
    max_length = max(len(original_word), len(entered_word))
    if max_length == 0:  # Handle edge case where both strings are empty
        return 5

    edit_distance = distance(original_word, entered_word)

    # Calculate similarity percentage
    similarity = (max_length - edit_distance) / max_length

    # Scale the similarity to a score between 0 and 5
    scaled_score = round(similarity * 5)

    return max(0, min(scaled_score, 5))  # Ensure score is within bounds
```

Output

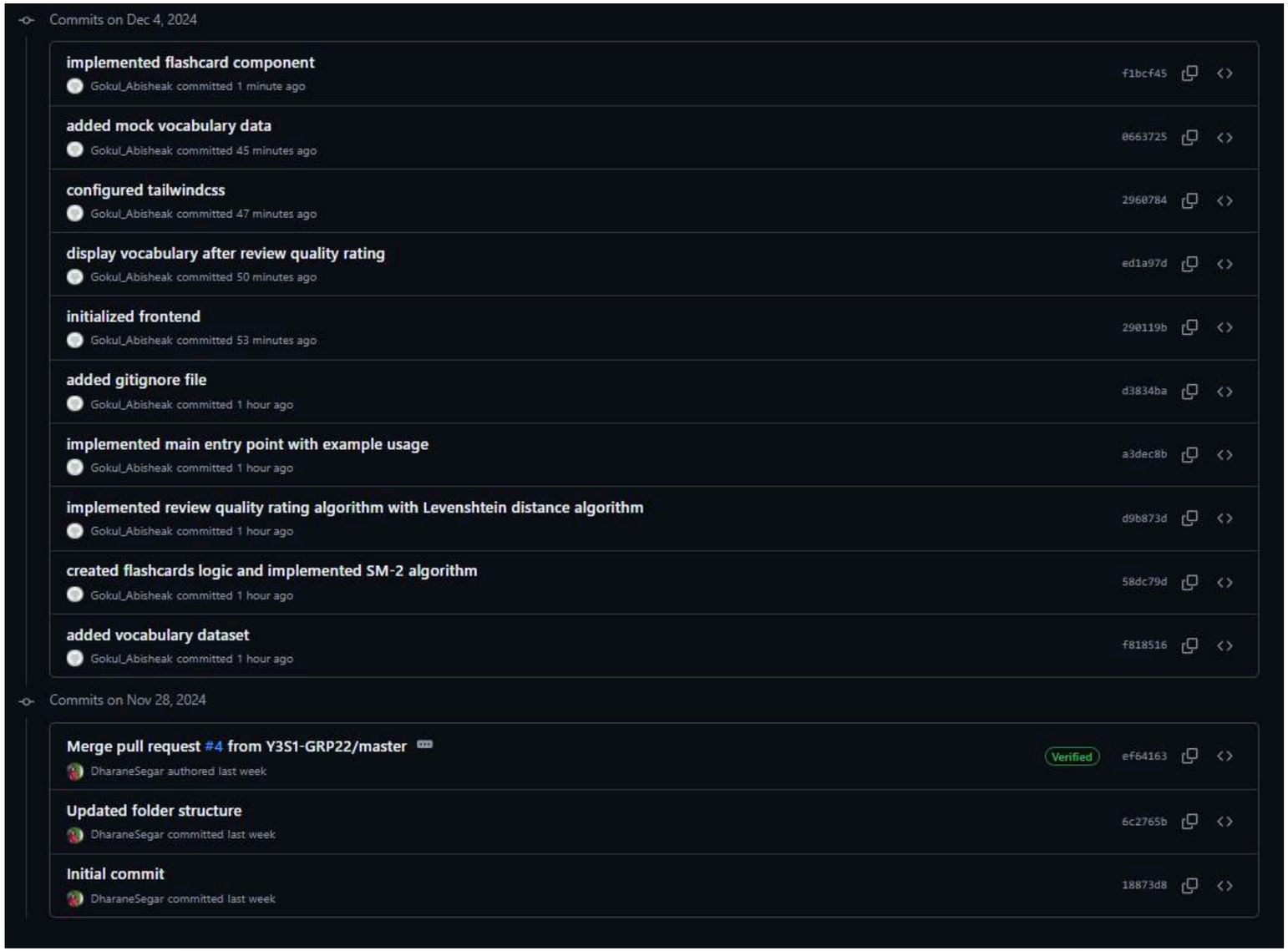
```
Day 1: Reviewing cards on 2024-12-04

Front: A living entity capable of growth, reproduction, and adaptation.
Enter your answer: organism
Calculated quality rating: 5

Front: The chemical processes occurring within a living organism.
Enter your answer: metabolism
Calculated quality rating: 5

Front: A group of atoms bonded together.
Enter your answer: molecules
Calculated quality rating: 4

Front: Specialized structures within a cell that perform specific functions.
Enter your answer: organ
Calculated quality rating: 3

Front: The basic unit of life in all living organisms.
Enter your answer: cell
Calculated quality rating: 5

Daily Review Log:
 - {'Card': 'A living entity capable of growth, reproduction, and adaptation.', 'Quality': 5, 'Next Review': '2024-12-05'}
 - {'Card': 'The chemical processes occurring within a living organism.', 'Quality': 5, 'Next Review': '2024-12-05'}
 - {'Card': 'A group of atoms bonded together.', 'Quality': 4, 'Next Review': '2024-12-05'}
 - {'Card': 'Specialized structures within a cell that perform specific functions.', 'Quality': 3, 'Next Review': '2024-12-05'}
 - {'Card': 'The basic unit of life in all living organisms.', 'Quality': 5, 'Next Review': '2024-12-05'}
```

```
Updated Card States:
Card: organism, Next Review: 2024-12-05, Repetitions: 1, Interval: 1.00, Easiness: 2.60
Card: metabolism, Next Review: 2024-12-05, Repetitions: 1, Interval: 1.00, Easiness: 2.60
Card: molecule, Next Review: 2024-12-05, Repetitions: 1, Interval: 1.00, Easiness: 2.50
Card: organelle, Next Review: 2024-12-05, Repetitions: 1, Interval: 1.00, Easiness: 2.36
Card: cell, Next Review: 2024-12-05, Repetitions: 1, Interval: 1.00, Easiness: 2.60
Card: tissue, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: organ, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: ecosystem, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: biosphere, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: carbohydrate, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: monosaccharide, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: disaccharide, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: polysaccharide, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: glycosidic, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: lipid, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: triglyceride, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: phospholipid, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: nucleoside, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: nucleotide, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
```

```
Updated Card States:
Card: organism, Next Review: 2024-12-11, Repetitions: 2, Interval: 6.00, Easiness: 2.60
Card: metabolism, Next Review: 2024-12-11, Repetitions: 2, Interval: 6.00, Easiness: 2.70
Card: molecule, Next Review: 2024-12-11, Repetitions: 2, Interval: 6.00, Easiness: 2.50
Card: organelle, Next Review: 2024-12-11, Repetitions: 2, Interval: 6.00, Easiness: 2.22
Card: cell, Next Review: 2024-12-11, Repetitions: 2, Interval: 6.00, Easiness: 2.70
Card: tissue, Next Review: 2024-12-12, Repetitions: 2, Interval: 6.00, Easiness: 2.60
Card: organ, Next Review: 2024-12-12, Repetitions: 2, Interval: 6.00, Easiness: 2.70
Card: ecosystem, Next Review: 2024-12-07, Repetitions: 1, Interval: 1.00, Easiness: 2.06
Card: biosphere, Next Review: 2024-12-07, Repetitions: 0, Interval: 1.00, Easiness: 1.30
Card: carbohydrate, Next Review: 2024-12-12, Repetitions: 2, Interval: 6.00, Easiness: 2.70
Card: monosaccharide, Next Review: 2024-12-07, Repetitions: 1, Interval: 1.00, Easiness: 2.50
Card: disaccharide, Next Review: 2024-12-07, Repetitions: 1, Interval: 1.00, Easiness: 2.60
Card: polysaccharide, Next Review: 2024-12-07, Repetitions: 1, Interval: 1.00, Easiness: 2.60
Card: glycosidic, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: lipid, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: triglyceride, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: phospholipid, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: nucleoside, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: nucleotide, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
Card: polynucleotide, Next Review: 2024-12-05, Repetitions: 0, Interval: 1.00, Easiness: 2.50
```

Flashcard UI

GitHub Commits



SLIIT FACULTY OF COMPUTING

IT21375132 |Srirajan G.A | 24-25J-257

Tasks to be done

- API development
- Adding gamification elements in frontend
- Integrating cognitive services in frontend
- Broaden vocabulary
- Backend integration

References

[1] J. Su, J. Ye, L. Nie, Y. Cao and Y. Chen, "Optimizing Spaced Repetition Schedule by Capturing the Dynamics of Memory," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 10, pp. 10085-10097, 1 Oct. 2023, doi: 10.1109/TKDE.2023.3251721.

[2] F. Schimanke, R. Mertens and B. S. Huck, "Retrieval of Relevant Data for Measuring the Impact of Spaced-Repetition Algorithms on the Learning Success in Mobile Learning Games," 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 2019, pp. 279-2795, doi: 10.1109/ISM46123.2019.00063.

[3] G. RANDAZZO, “Memory models for spaced repetition systems,” Polimi.it, Sep. 2023, doi: http://hdl.handle.net/10589/186407.

[4] F. Schimanke, "The Impact of Spaced Repetition Learning on the Learning Success in Mobile Learning Games," 2021 IEEE International Symposium on Multimedia (ISM), Naple, Italy, 2021, pp. 275-280, doi: 10.1109/ISM52913.2021.00054.

IT21068478

Dharane.S
Software Engineering

LLM BASED ABSTRACTIVE TEXT SUMMARIZATION TOOL WITH VOICE OUTPUT IMPLEMENTED IN DIFFERENT SOFTWARE ARCHITECTURES

Introduction

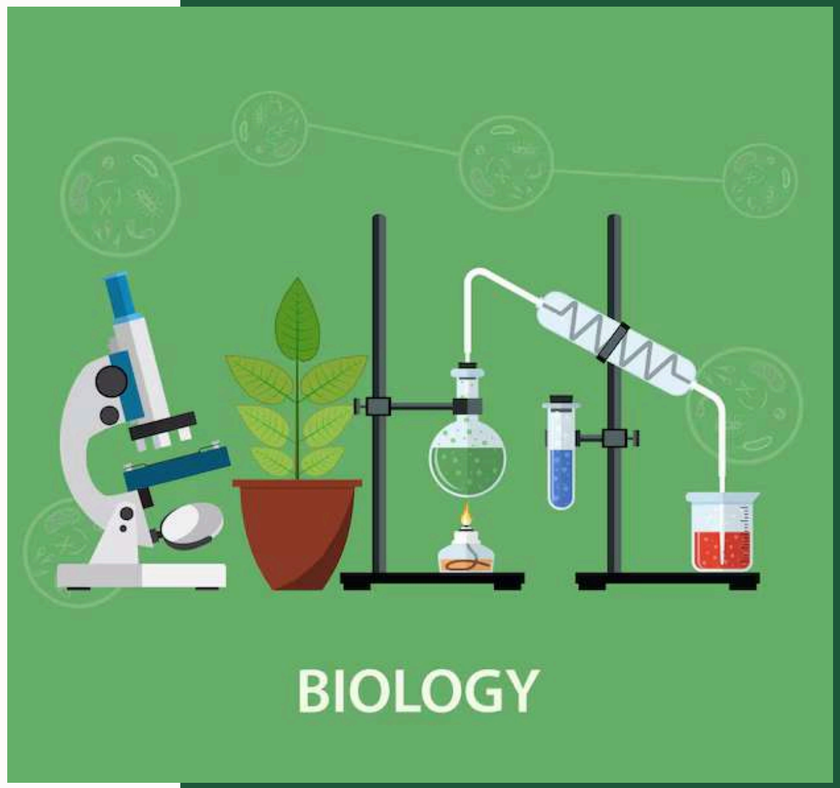
01 Background

02 Research Problem

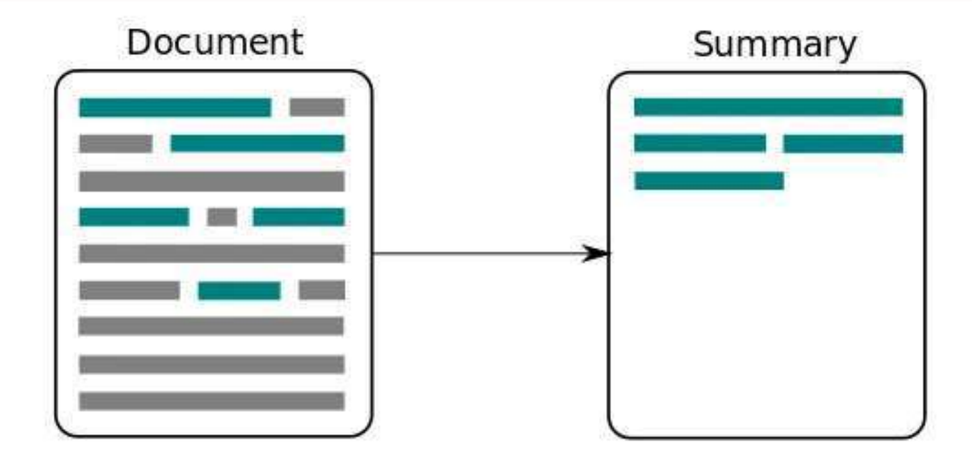
03 Research Gap

04 Main and Sub Objectives

05 Methodology



BACKGROUND



What is summarization?

What is NLP?

RESEARCH PROBLEM

01

How can text summarization be optimized to provide concise, exam-focused summaries for A/L biology students?

How can the integration of voice output features with customizable word counts in an e-learning summarization tool enhance accessibility, catering to different learning preferences and needs?

02

Research Gap

	Document upload	Customizable word count	Audible summary	Extract data from approved resources
grammarly	✗	✗	✗	✗
QuillBot	✓	✗	✗	✗
BIOMENTOR	✓	✓	✓	✓

OBJECTIVES

Objective 1

Ensure Accuracy: Use government-approved resources to maintain content accuracy and educational standards.

Objective 3

User customization: Implement customizable word count settings for summaries.

Main Objective

Create a tool that extracts and summarizes key concepts from A/L biology resources based on a given topic, and generates concise summaries from uploaded text documents, specifically to aid students preparing for A/L exams.

Objective 2

Enhance Accessibility: Integrate a voice output feature to provide audible summaries.

Objective 4

Various implementation: Implement the component in various architectures, compare, and analyze their strengths and weaknesses.

IT21068478 |Dharane. S | 24-25J-257

Methodology

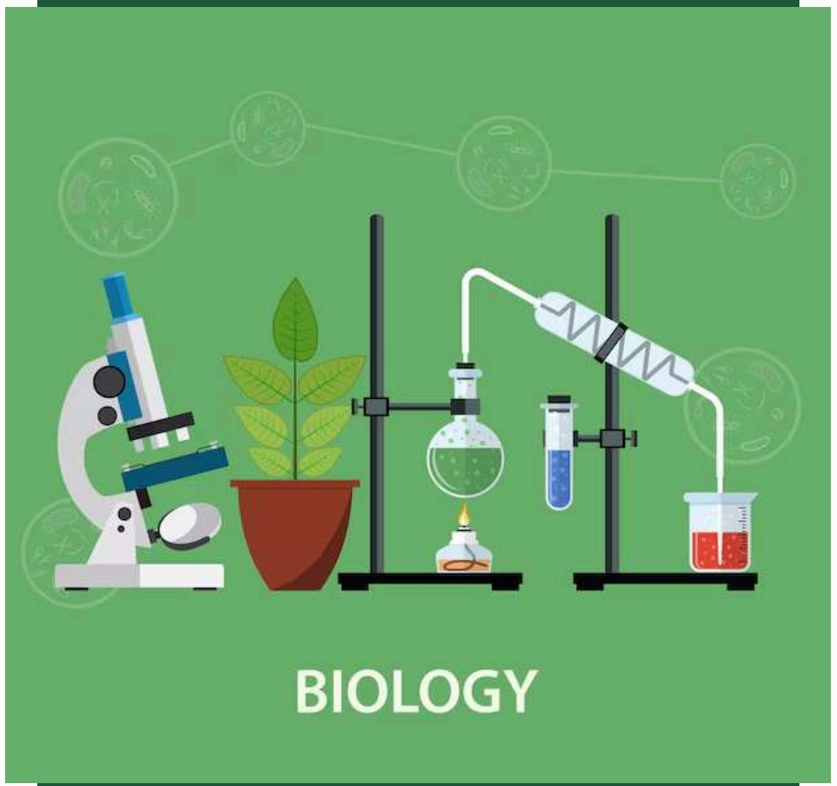
01 System Diagram

02 Tools and Technologies

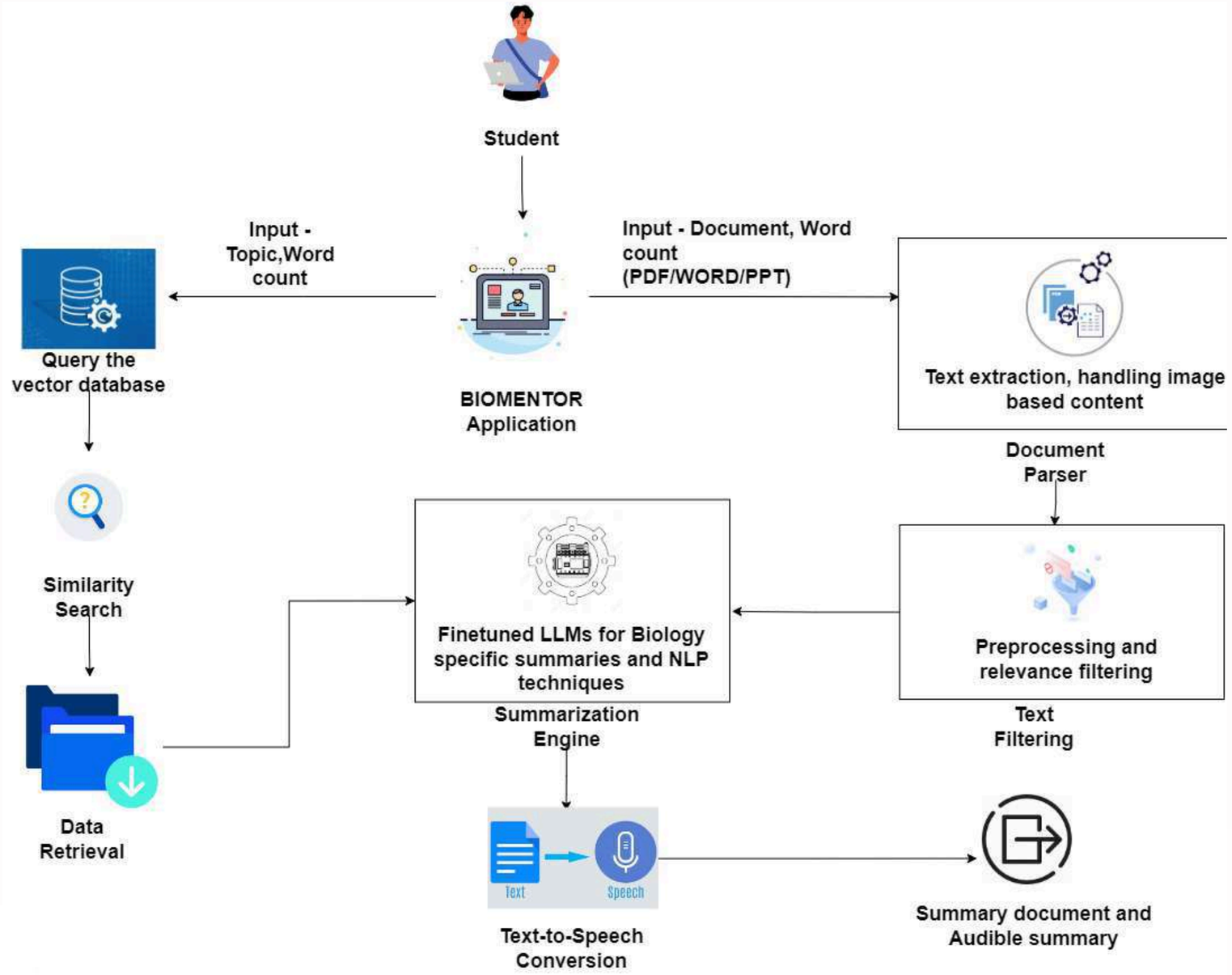
03 Requirements

04 Work Breakdown Structure

05 Gantt Chart



System Diagram



Tools & Technologies

Programming Languages

Python
React Js

Database

Faiss
Mongo DB

Project Management

Jira

Other tools

Git
Draw.io
Postman

Frameworks

Transformer model
pandas
Pytorch
pyttsx3/gtts
OCR
Flask
numpy

SLIIT
FACULTY OF COMPUTING

IT21068478 |Dharane. S | 24-25J-257

Requirements

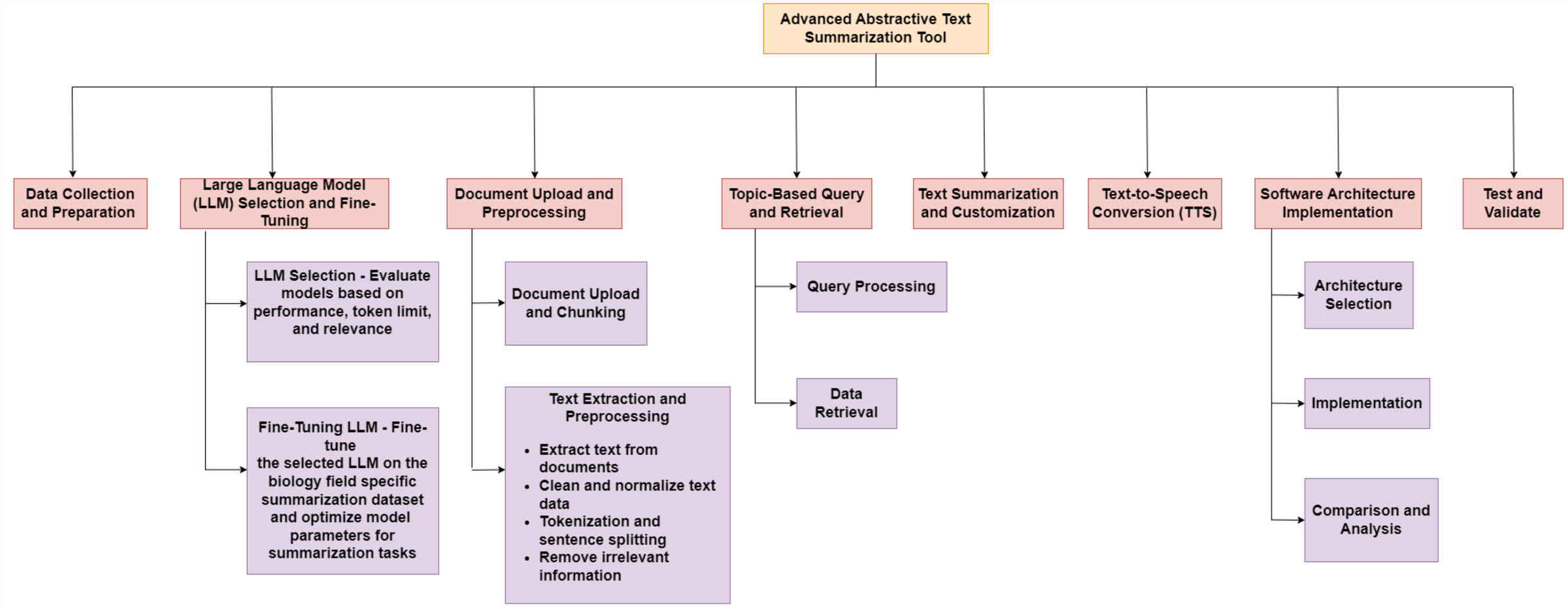
Functional

1. The tool should accept any type of text documents, and extract text from image-based documents using OCR.
2. Collect data from government-approved A/L Biology resources.
3. The tool should extract relevant text from documents.
4. The tool generates concise summaries based on user-defined word counts.
5. The tool should convert summarized text into voice output.

Non-Functional

1. Accuracy
2. Performance
3. Availability
4. Compliance
5. Usability

Work Breakdown Structure



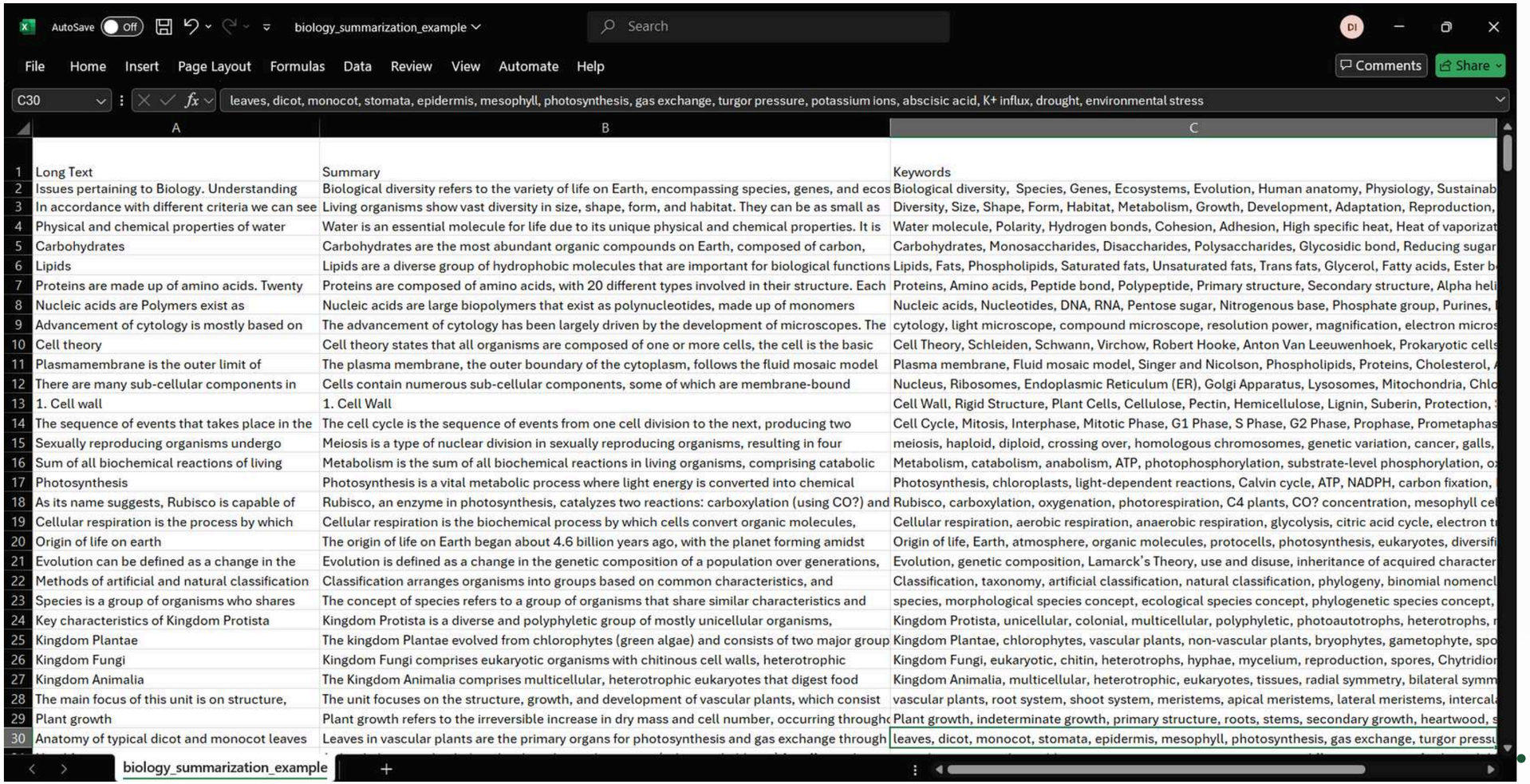
Completion of the project

SLIIT
FACULTY OF COMPUTING

IT21068478 |Dharane. S | 24-25J-257

Data Collection

Summarization Dataset

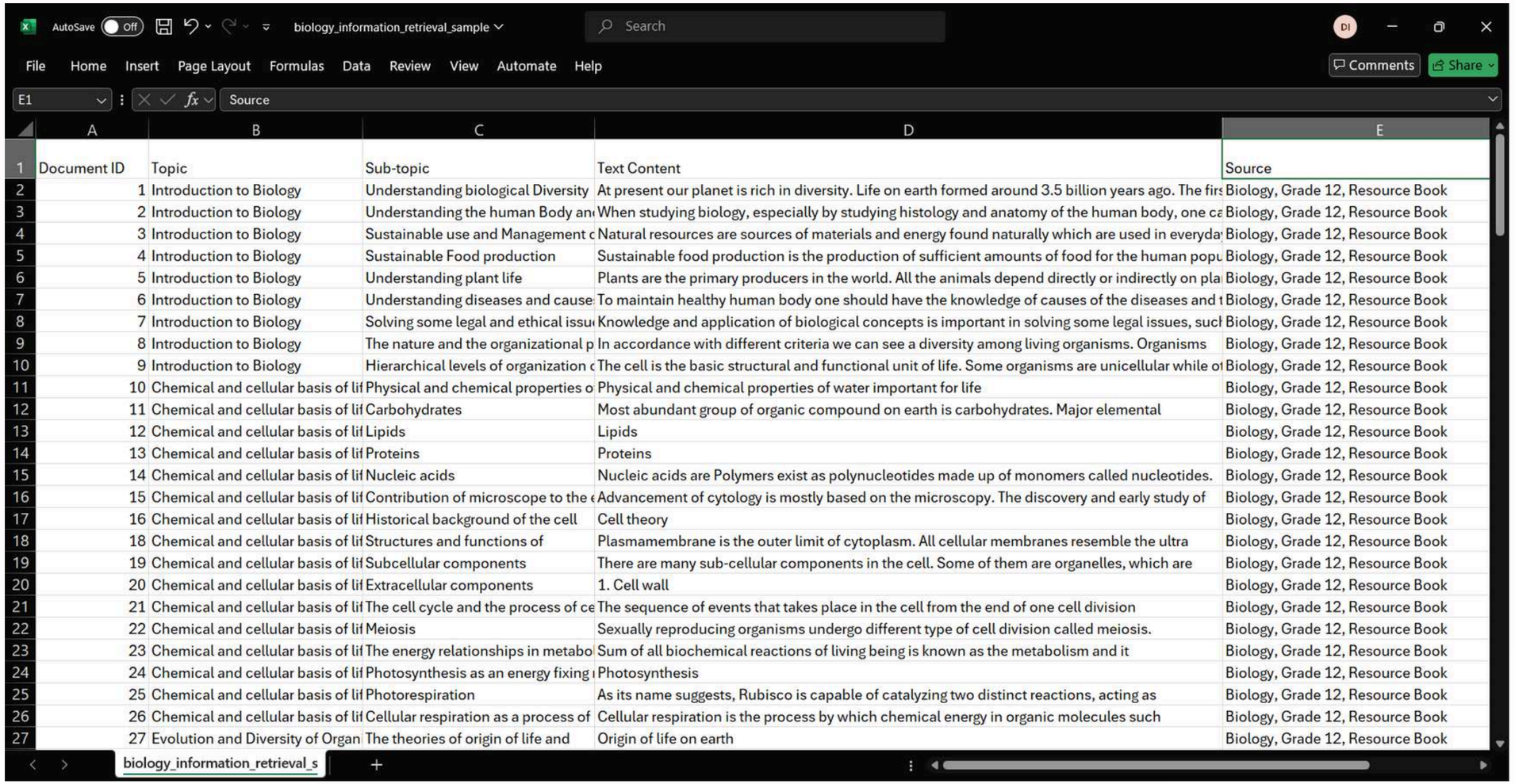


biology_summarization_example

	A	B	C
1	Long Text	Summary	Keywords
2	Issues pertaining to Biology. Understanding	Biological diversity refers to the variety of life on Earth, encompassing species, genes, and ecos	Biological diversity, Species, Genes, Ecosystems, Evolution, Human anatomy, Physiology, Sustainab
3	In accordance with different criteria we can see	Living organisms show vast diversity in size, shape, form, and habitat. They can be as small as	Diversity, Size, Shape, Form, Habitat, Metabolism, Growth, Development, Adaptation, Reproduction,
4	Physical and chemical properties of water	Water is an essential molecule for life due to its unique physical and chemical properties. It is	Water molecule, Polarity, Hydrogen bonds, Cohesion, Adhesion, High specific heat, Heat of vaporizat
5	Carbohydrates	Carbohydrates are the most abundant organic compounds on Earth, composed of carbon,	Carbohydrates, Monosaccharides, Disaccharides, Polysaccharides, Glycosidic bond, Reducing sugar
6	Lipids	Lipids are a diverse group of hydrophobic molecules that are important for biological functions	Lipids, Fats, Phospholipids, Saturated fats, Unsaturated fats, Trans fats, Glycerol, Fatty acids, Ester b
7	Proteins are made up of amino acids. Twenty	Proteins are composed of amino acids, with 20 different types involved in their structure. Each	Proteins, Amino acids, Peptide bond, Polypeptide, Primary structure, Secondary structure, Alpha heli
8	Nucleic acids are Polymers exist as	Nucleic acids are large biopolymers that exist as polynucleotides, made up of monomers	Nucleic acids, Nucleotides, DNA, RNA, Pentose sugar, Nitrogenous base, Phosphate group, Purines,
9	Advancement of cytology is mostly based on	The advancement of cytology has been largely driven by the development of microscopes. The	cytology, light microscope, compound microscope, resolution power, magnification, electron micros
10	Cell theory	Cell theory states that all organisms are composed of one or more cells, the cell is the basic	Cell Theory, Schleiden, Schwann, Virchow, Robert Hooke, Anton Van Leeuwenhoek, Prokaryotic cells
11	Plasmamembrane is the outer limit of	The plasma membrane, the outer boundary of the cytoplasm, follows the fluid mosaic model	Plasma membrane, Fluid mosaic model, Singer and Nicolson, Phospholipids, Proteins, Cholesterol,
12	There are many sub-cellular components in	Cells contain numerous sub-cellular components, some of which are membrane-bound	Nucleus, Ribosomes, Endoplasmic Reticulum (ER), Golgi Apparatus, Lysosomes, Mitochondria, Chlo
13	1. Cell wall	1. Cell Wall	Cell Wall, Rigid Structure, Plant Cells, Cellulose, Pectin, Hemicellulose, Lignin, Suberin, Protection,
14	The sequence of events that takes place in the	The cell cycle is the sequence of events from one cell division to the next, producing two	Cell Cycle, Mitosis, Interphase, Mitotic Phase, G1 Phase, S Phase, G2 Phase, Prophase, Prometaphas
15	Sexually reproducing organisms undergo	Meiosis is a type of nuclear division in sexually reproducing organisms, resulting in four	meiosis, haploid, diploid, crossing over, homologous chromosomes, genetic variation, cancer, galls,
16	Sum of all biochemical reactions of living	Metabolism is the sum of all biochemical reactions in living organisms, comprising catabolic	Metabolism, catabolism, anabolism, ATP, photophosphorylation, substrate-level phosphorylation, o
17	Photosynthesis	Photosynthesis is a vital metabolic process where light energy is converted into chemical	Photosynthesis, chloroplasts, light-dependent reactions, Calvin cycle, ATP, NADPH, carbon fixation,
18	As its name suggests, Rubisco is capable of	Rubisco, an enzyme in photosynthesis, catalyzes two reactions: carboxylation (using CO?) and	Rubisco, carboxylation, oxygenation, photorespiration, C4 plants, CO? concentration, mesophyll ce
19	Cellular respiration is the process by which	Cellular respiration is the biochemical process by which cells convert organic molecules,	Cellular respiration, aerobic respiration, anaerobic respiration, glycolysis, citric acid cycle, electron t
20	Origin of life on earth	The origin of life on Earth began about 4.6 billion years ago, with the planet forming amidst	Origin of life, Earth, atmosphere, organic molecules, protocells, photosynthesis, eukaryotes, diversifi
21	Evolution can be defined as a change in the	Evolution is defined as a change in the genetic composition of a population over generations,	Evolution, genetic composition, Lamarck's Theory, use and disuse, inheritance of acquired character
22	Methods of artificial and natural classification	Classification arranges organisms into groups based on common characteristics, and	Classification, taxonomy, artificial classification, natural classification, phylogeny, binomial nomencl
23	Species is a group of organisms who shares	The concept of species refers to a group of organisms that share similar characteristics and	species, morphological species concept, ecological species concept, phylogenetic species concept,
24	Key characteristics of Kingdom Protista	Kingdom Protista is a diverse and polyphyletic group of mostly unicellular organisms,	Kingdom Protista, unicellular, colonial, multicellular, polyphyletic, photoautotrophs, heterotrophs, r
25	Kingdom Plantae	The kingdom Plantae evolved from chlorophytes (green algae) and consists of two major group	Kingdom Plantae, chlorophytes, vascular plants, non-vascular plants, bryophytes, gametophyte, spo
26	Kingdom Fungi	Kingdom Fungi comprises eukaryotic organisms with chitinous cell walls, heterotrophic	Kingdom Fungi, eukaryotic, chitin, heterotrophs, hyphae, mycelium, reproduction, spores, Chytridior
27	Kingdom Animalia	The Kingdom Animalia comprises multicellular, heterotrophic eukaryotes that digest food	Kingdom Animalia, multicellular, heterotrophic, eukaryotes, tissues, radial symmetry, bilateral symm
28	The main focus of this unit is on structure,	The unit focuses on the structure, growth, and development of vascular plants, which consist	vascular plants, root system, shoot system, meristems, apical meristems, lateral meristems, intercal
29	Plant growth	Plant growth refers to the irreversible increase in dry mass and cell number, occurring through	Plant growth, indeterminate growth, primary structure, roots, stems, secondary growth, heartwood, s
30	Anatomy of typical dicot and monocot leaves	Leaves in vascular plants are the primary organs for photosynthesis and gas exchange through	leaves, dicot, monocot, stomata, epidermis, mesophyll, photosynthesis, gas exchange, turgor pressu

Data Collection

Information Retrieval Dataset



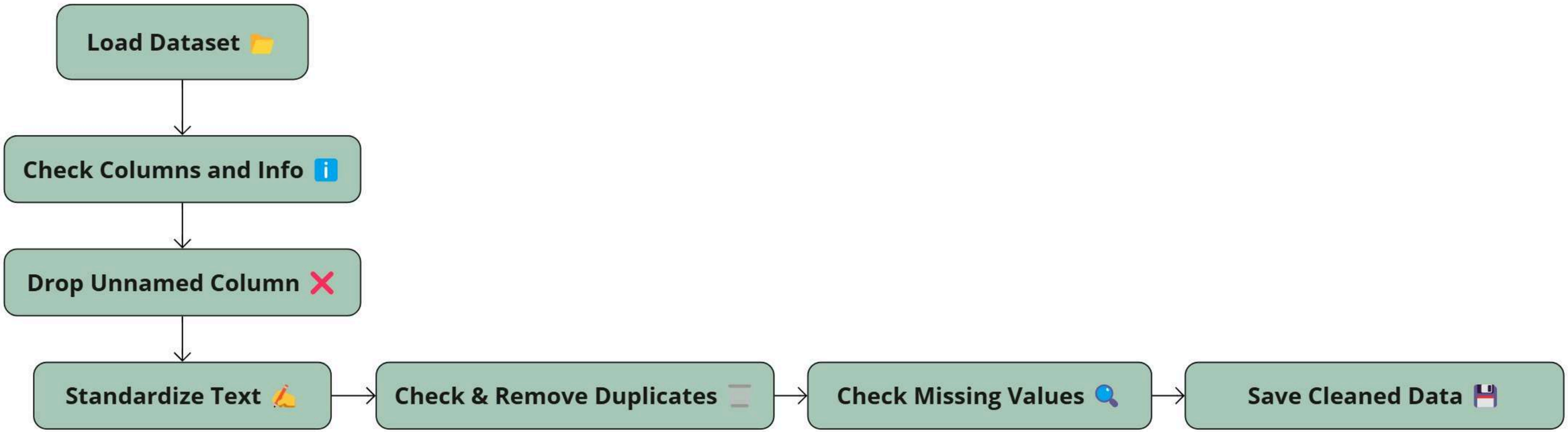
	A	B	C	D	E
1	Document ID	Topic	Sub-topic	Text Content	Source
2	1	Introduction to Biology	Understanding biological Diversity	At present our planet is rich in diversity. Life on earth formed around 3.5 billion years ago. The firs	Biology, Grade 12, Resource Book
3	2	Introduction to Biology	Understanding the human Body an	When studying biology, especially by studying histology and anatomy of the human body, one ca	Biology, Grade 12, Resource Book
4	3	Introduction to Biology	Sustainable use and Management c	Natural resources are sources of materials and energy found naturally which are used in everyda	Biology, Grade 12, Resource Book
5	4	Introduction to Biology	Sustainable Food production	Sustainable food production is the production of sufficient amounts of food for the human popu	Biology, Grade 12, Resource Book
6	5	Introduction to Biology	Understanding plant life	Plants are the primary producers in the world. All the animals depend directly or indirectly on pla	Biology, Grade 12, Resource Book
7	6	Introduction to Biology	Understanding diseases and cause	To maintain healthy human body one should have the knowledge of causes of the diseases and t	Biology, Grade 12, Resource Book
8	7	Introduction to Biology	Solving some legal and ethical issu	Knowledge and application of biological concepts is important in solving some legal issues, such	Biology, Grade 12, Resource Book
9	8	Introduction to Biology	The nature and the organizational p	In accordance with different criteria we can see a diversity among living organisms. Organisms	Biology, Grade 12, Resource Book
10	9	Introduction to Biology	Hierarchical levels of organization c	The cell is the basic structural and functional unit of life. Some organisms are unicellular while ot	Biology, Grade 12, Resource Book
11	10	Chemical and cellular basis of lif	Physical and chemical properties o	Physical and chemical properties of water important for life	Biology, Grade 12, Resource Book
12	11	Chemical and cellular basis of lif	Carbohydrates	Most abundant group of organic compound on earth is carbohydrates. Major elemental	Biology, Grade 12, Resource Book
13	12	Chemical and cellular basis of lif	Lipids	Lipids	Biology, Grade 12, Resource Book
14	13	Chemical and cellular basis of lif	Proteins	Proteins	Biology, Grade 12, Resource Book
15	14	Chemical and cellular basis of lif	Nucleic acids	Nucleic acids are Polymers exist as polynucleotides made up of monomers called nucleotides.	Biology, Grade 12, Resource Book
16	15	Chemical and cellular basis of lif	Contribution of microscope to the c	Advancement of cytology is mostly based on the microscopy. The discovery and early study of	Biology, Grade 12, Resource Book
17	16	Chemical and cellular basis of lif	Historical background of the cell	Cell theory	Biology, Grade 12, Resource Book
18	18	Chemical and cellular basis of lif	Structures and functions of	Plasmamembrane is the outer limit of cytoplasm. All cellular membranes resemble the ultra	Biology, Grade 12, Resource Book
19	19	Chemical and cellular basis of lif	Subcellular components	There are many sub-cellular components in the cell. Some of them are organelles, which are	Biology, Grade 12, Resource Book
20	20	Chemical and cellular basis of lif	Extracellular components	1. Cell wall	Biology, Grade 12, Resource Book
21	21	Chemical and cellular basis of lif	The cell cycle and the process of ce	The sequence of events that takes place in the cell from the end of one cell division	Biology, Grade 12, Resource Book
22	22	Chemical and cellular basis of lif	Meiosis	Sexually reproducing organisms undergo different type of cell division called meiosis.	Biology, Grade 12, Resource Book
23	23	Chemical and cellular basis of lif	The energy relationships in metabo	Sum of all biochemical reactions of living being is known as the metabolism and it	Biology, Grade 12, Resource Book
24	24	Chemical and cellular basis of lif	Photosynthesis as an energy fixing r	Photosynthesis	Biology, Grade 12, Resource Book
25	25	Chemical and cellular basis of lif	Photorespiration	As its name suggests, Rubisco is capable of catalyzing two distinct reactions, acting as	Biology, Grade 12, Resource Book
26	26	Chemical and cellular basis of lif	Cellular respiration as a process of	Cellular respiration is the process by which chemical energy in organic molecules such	Biology, Grade 12, Resource Book
27	27	Evolution and Diversity of Organ	The theories of origin of life and	Origin of life on earth	Biology, Grade 12, Resource Book

Model Selection

Why Flan-T5?

- Instruction-Following Capability
- Versatility in Fine-tuning
- Efficiency and Scalability
- Customizability
- Pre-trained Knowledge
- Resource Compatibility

Data pre-processing



Data pre-processing

DataPreprocessing.ipynb

DataPreprocessing.ipynb > # Step 3: Inspect the dataset

+ Code + Markdown | Run All Clear All Outputs | Outline ···

```
#Removing unwanted characters, space, special characters
import re

# Function to standardize text: convert to lowercase, remove irrelevant symbols
def standardize_text(text):
    # Convert to lowercase
    text = text.lower()
    # Remove irrelevant symbols (e.g., unusual bullet points, etc.)
    text = re.sub(r'[¦]', '', text)
    # Remove any remaining special characters (e.g., non-alphanumeric symbols)
    text = re.sub(r'[^a-zA-Z0-9\s]', '', text)
    # Remove extra whitespace
    text = re.sub(r'\s+', ' ', text).strip()
    return text

# Apply the function to standardize the 'Long Text' and 'Summary' columns
data['Long Text'] = data['Long Text'].apply(standardize_text)
data['Summary'] = data['Summary'].apply(standardize_text)

# Display a sample to verify the standardization process
data[['Long Text', 'Summary']].head()
```

	Long Text	Summary
0	issues pertaining to biology understanding bio...	biological diversity refers to the variety of ...
1	in accordance with different criteria we can s...	living organisms show vast diversity in size s...
2	physical and chemical properties of water impo...	water is an essential molecule for life due to...
3	carbohydrates most abundant group of organic c...	carbohydrates are the most abundant organic co...
4	lipids diverse group of hydrophobic molecules ...	lipids are a diverse group of hydrophobic mole...

```
# Checking and removing duplicates based on specific columns 'Long Text' and 'Summary' only, which are the main content columns
data_cleaned = data.drop_duplicates(subset=['Long Text', 'Summary'])

# Display the number of rows before and after removing duplicates
original_count = data.shape[0]
cleaned_count = data_cleaned.shape[0]

original_count, cleaned_count
```

[18]

DataPreprocessing.ipynb

DataPreprocessing.ipynb > # Step 3: Inspect the dataset

+ Code + Markdown | Run All Clear All Outputs | Outline ···

```
# Checking for missing values in the essential columns 'Long Text' and 'Summary'
missing_long_text = data['Long Text'].isnull().sum()
missing_summary = data['Summary'].isnull().sum()

# Displaying the count of missing values in each essential column
missing_long_text, missing_summary
```

[19]

```
(0, 0)
```

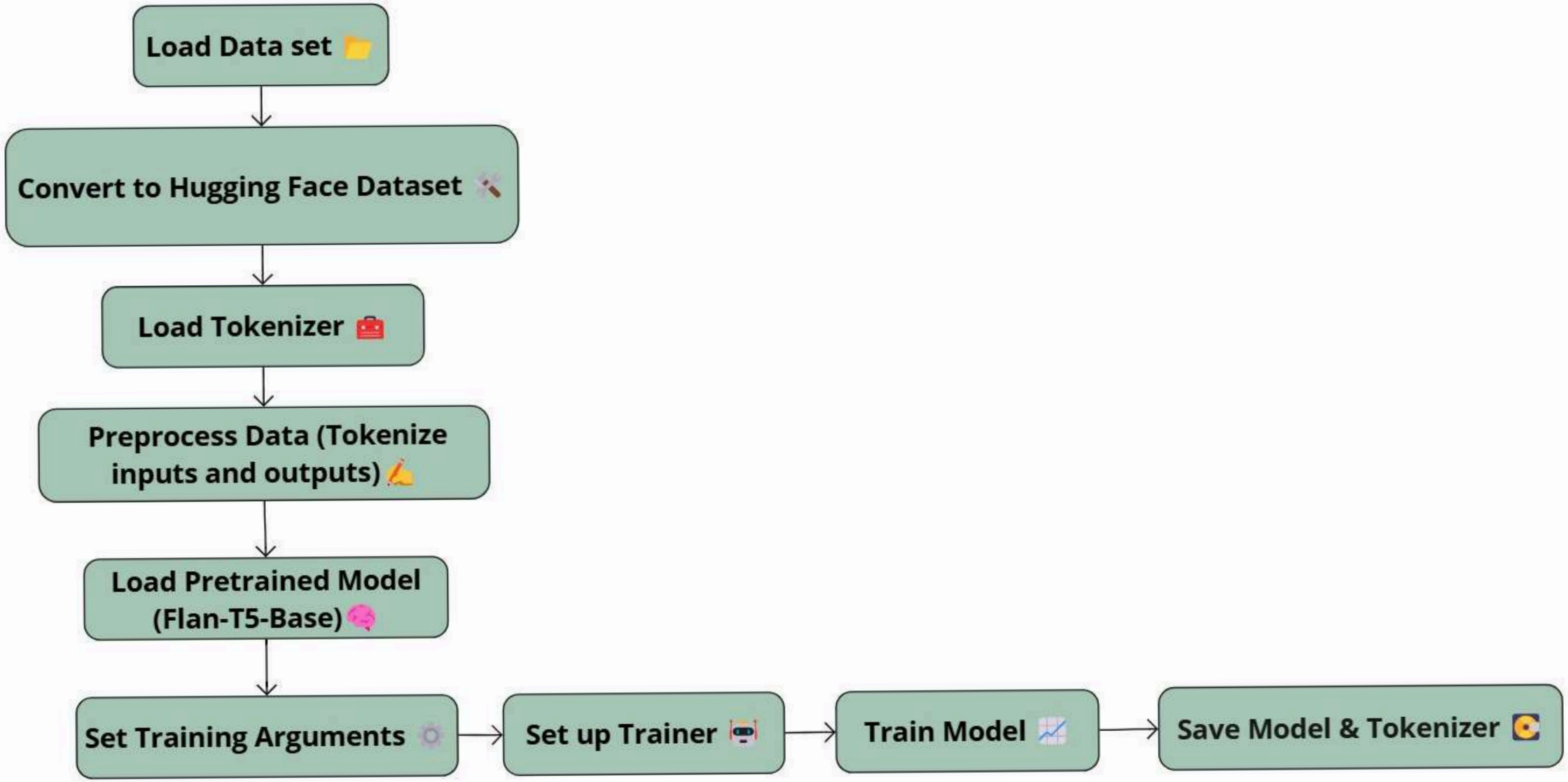
```
# Saving the cleaned dataset to a new CSV file named 'bio_summary.csv'
output_path = 'D:/Downloads/RP/Summarization/Summary_Description/bio_summary.csv'
data_cleaned.to_csv(output_path, index=False)

output_path
```

```
'C:/Users/dinon/Desktop/Summary/bio_summary.csv'
```

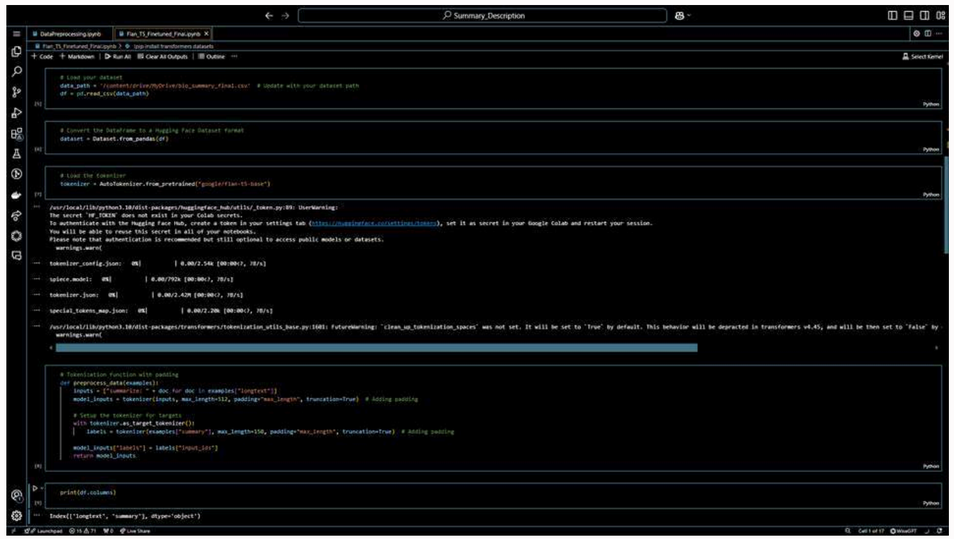
```
# Selecting only 'Long Text' and 'Summary' columns and renaming the headers to lowercase
data_final = data_cleaned[['Long Text', 'Summary']].rename(columns={'Long Text': 'long text', 'Summary': 'summary'})

# Saving this final cleaned dataset to a new CSV file
output_path_final = 'D:/Downloads/RP/Summarization/Summary_Description/bio_summary_final.csv'
data_final.to_csv(output_path_final, index=False)

output_path_final
```

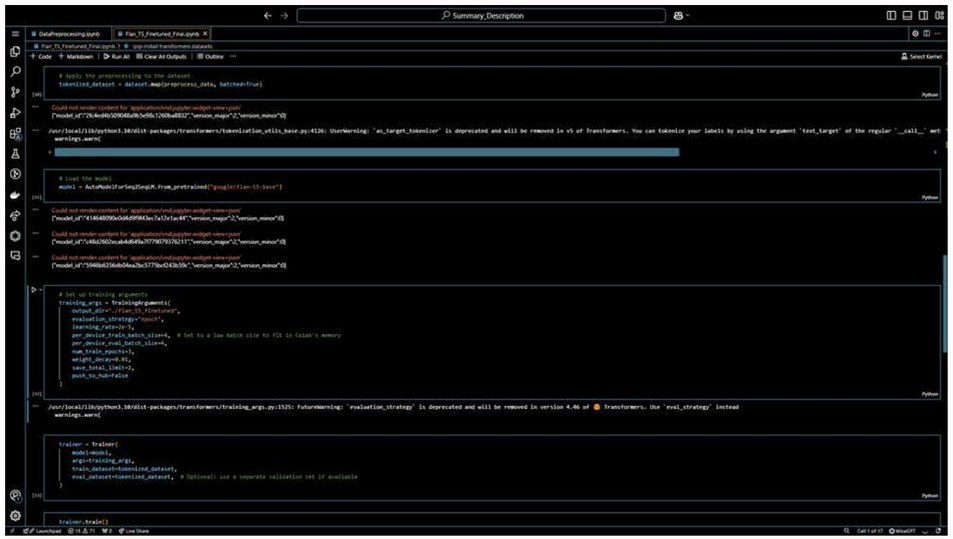
```
'C:/Users/dinon/Desktop/Summary/bio_summary_final.csv'
```

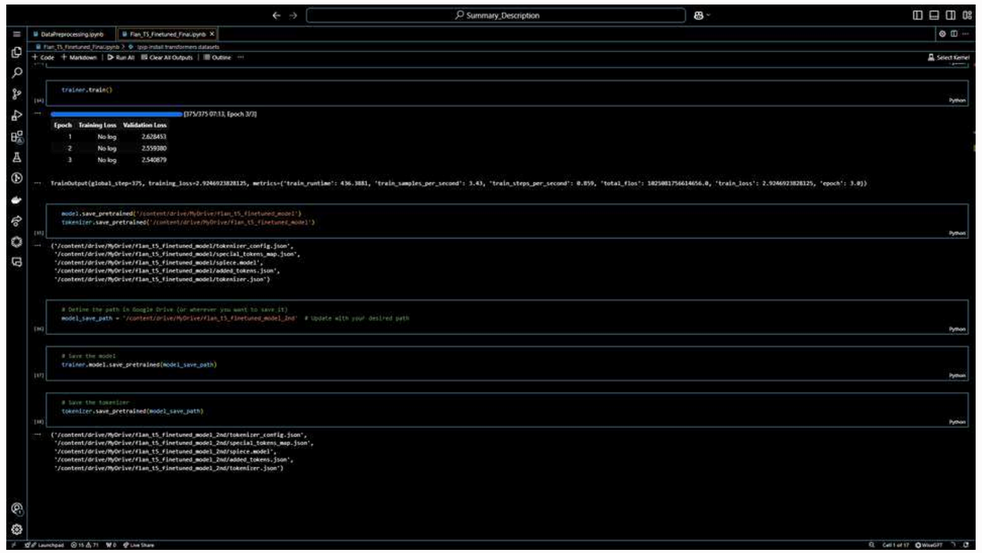
Fine-tuning the LLM



Fine-tuning the LLM







Evaluate the Fine-tuned model

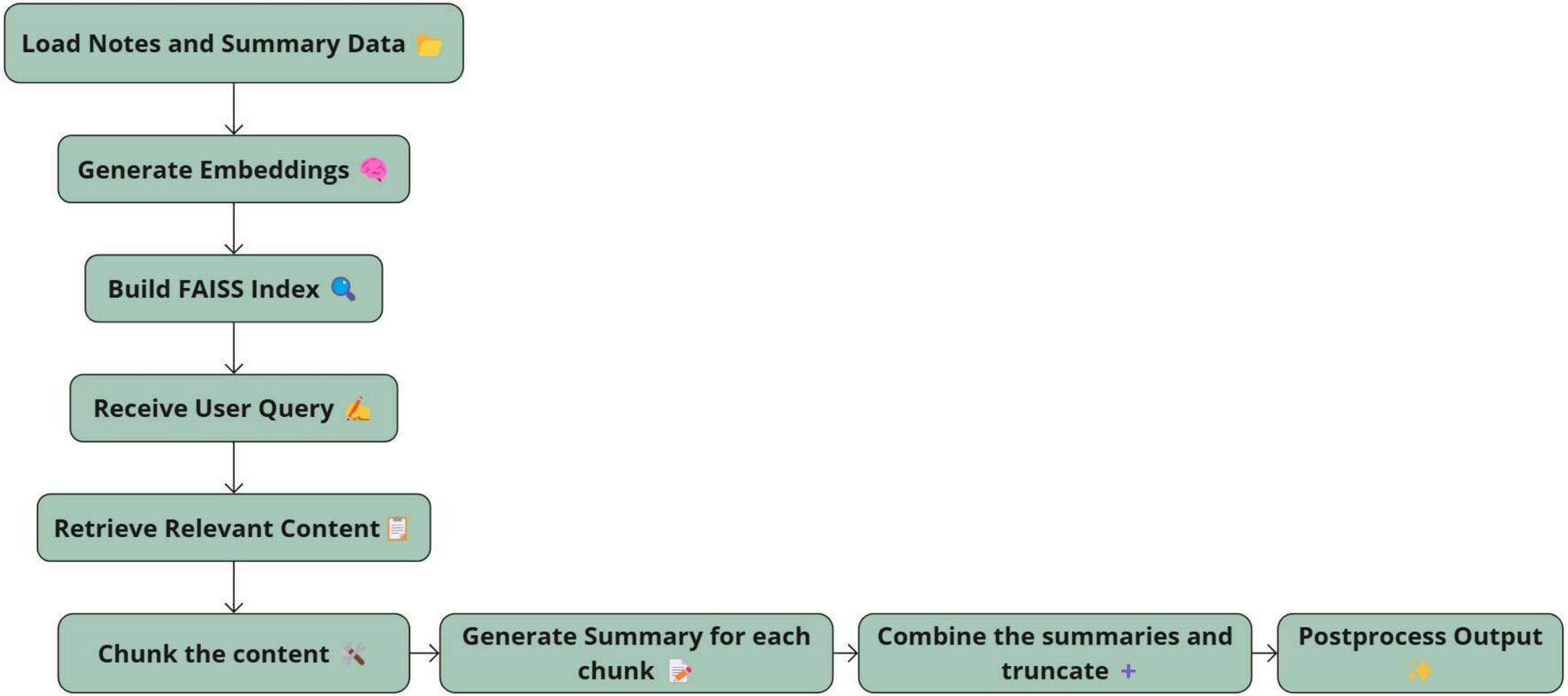
Evaluation Metric: ROUGE(Recall-Oriented Understudy for Gisting Evaluation)

- Measures word overlap between the generated summary and the reference summary.
- Types of ROUGE Scores
  ROUGE-1: Unigram (single word) overlap.
  ROUGE-2: Bigram (two-word) overlap.
  ROUGE-L: Longest Common Subsequence (LCS) match.

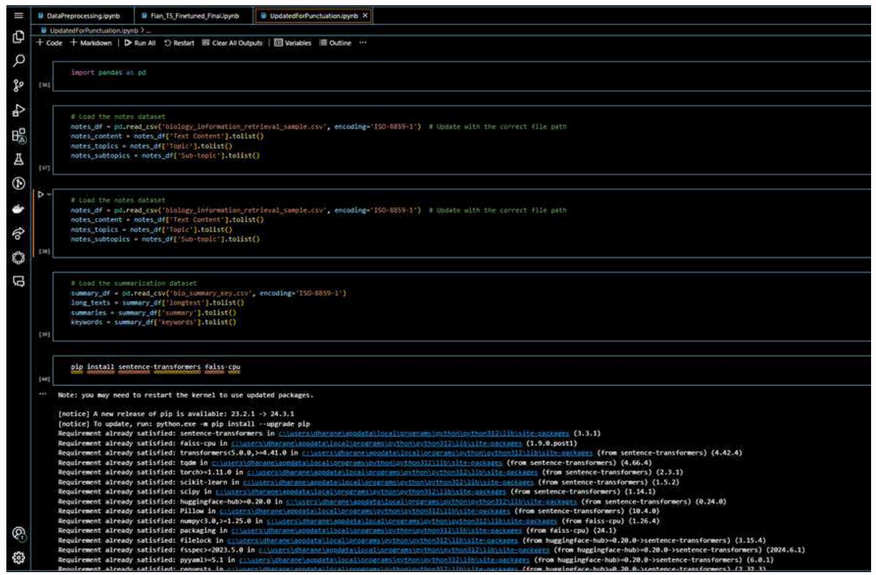
Evaluate the Fine-tuned model

{'**rouge1**': Score(**precision=0.7414965986394558**, recall=0.5989010989010989, fmeasure=0.662613981762917 9), '**rouge2**': Score(**precision=0.4041095890410959**, recall=0.3259668508287293, fmeasure=0.3608562691131499), '**rougeL**': Score(**precision=0.5442176870748 3**, recall=0.4395604395604395 5, fmeasure=0.4863221884498480 4)}

RAG Implementation

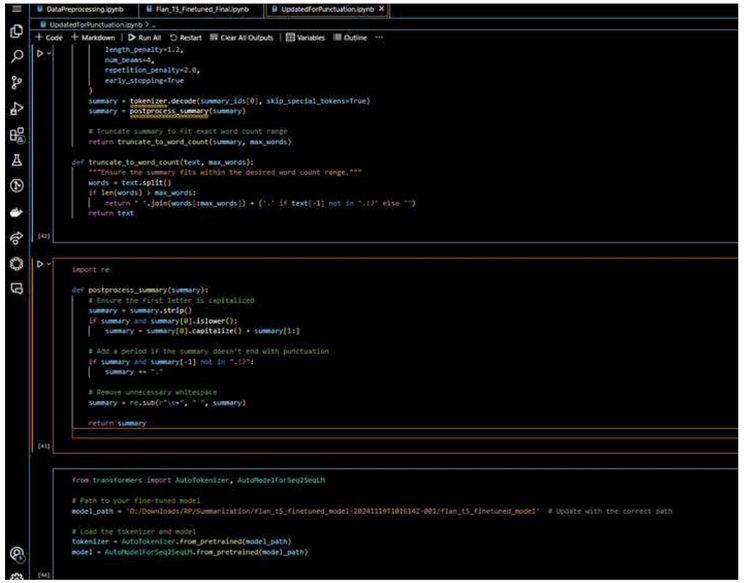


RAG Implementation



```
import pandas as pd

# Load the notes dataset
notes_df = pd.read_csv('Biology_information_retrieval_sample.csv', encoding='ISO-8859-1')  # Update with the correct file path
notes_content = notes_df['Text Content'].tolist()
notes_topics = notes_df['Topic'].tolist()
notes_subtopics = notes_df['Sub-topic'].tolist()

# Load the notes dataset
notes_df = pd.read_csv('Biology_information_retrieval_sample.csv', encoding='ISO-8859-1')  # Update with the correct file path
notes_content = notes_df['Text Content'].tolist()
notes_topics = notes_df['Topic'].tolist()
notes_subtopics = notes_df['Sub-topic'].tolist()

# Load the summarization dataset
summary_df = pd.read_csv('bio_summary_key.csv', encoding='ISO-8859-1')
long_texts = summary_df['longtext'].tolist()
summaries = summary_df['summary'].tolist()
keywords = summary_df['keywords'].tolist()

pip install sentence-transformers faiss-cpu
```

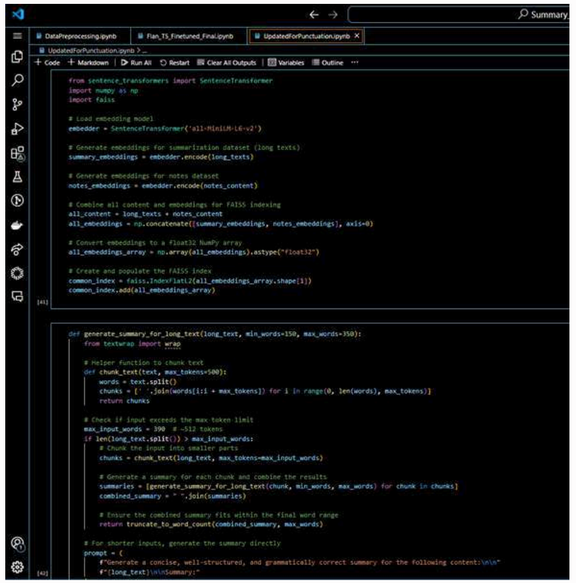
```
Note: you may need to restart the kernel to use updated packages.

[notice] A new release of pip is available: 23.2.1 -> 24.3.1
[notice] To update, run: python.exe -m pip install --upgrade pip
Requirement already satisfied: sentence-transformers in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (3.3.1)
Requirement already satisfied: faiss-cpu in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (1.9.0.post1)
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from sentence-transformers) (4.42.4)
Requirement already satisfied: tqdm in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from sentence-transformers) (4.66.4)
Requirement already satisfied: torch>=1.11.0 in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from sentence-transformers) (2.3.1)
Requirement already satisfied: scikit-learn in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from sentence-transformers) (1.5.2)
Requirement already satisfied: scipy in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from sentence-transformers) (1.14.1)
Requirement already satisfied: huggingface-hub>=0.20.0 in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from sentence-transformers) (0.24.0)
Requirement already satisfied: Pillow in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from sentence-transformers) (10.4.0)
Requirement already satisfied: numpy<3.0,>=1.25.0 in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from faiss-cpu) (1.26.4)
Requirement already satisfied: packaging in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from faiss-cpu) (24.1)
Requirement already satisfied: filelock in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from huggingface-hub>=0.20.0->sentence-transformers) (3.15.4)
Requirement already satisfied: fsspec>=2023.5.0 in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from huggingface-hub>=0.20.0->sentence-transformers) (2024.6.1)
Requirement already satisfied: pyyaml>=5.1 in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from huggingface-hub>=0.20.0->sentence-transformers) (6.0.1)
Requirement already satisfied: requests in c:\users\dharane\appdata\local\programs\python\python312\lib\site-packages (from huggingface-hub>=0.20.0->sentence-transformers) (2.32.3)
```



```
        length_penalty=1.2,
        num_beams=4,
        repetition_penalty=2.0,
        early_stopping=True
    )
    summary = tokenizer.decode(summary_ids[0], skip_special_tokens=True)
    summary = postprocess_summary(summary)

    # Truncate summary to fit exact word count range
    return truncate_to_word_count(summary, max_words)

def truncate_to_word_count(text, max_words):
    """Ensure the summary fits within the desired word count range."""
    words = text.split()
    if len(words) > max_words:
        return " ".join(words[:max_words]) + ('.' if text[-1] not in '.!?' else '')
    return text
```

```
import re

def postprocess_summary(summary):
    # Ensure the first letter is capitalized
    summary = summary.strip()
    if summary and summary[0].islower():
        summary = summary[0].capitalize() + summary[1:]

    # Add a period if the summary doesn't end with punctuation
    if summary and summary[-1] not in ".!?":
        summary += "."

    # Remove unnecessary whitespace
    summary = re.sub(r"\s+", " ", summary)

    return summary
```

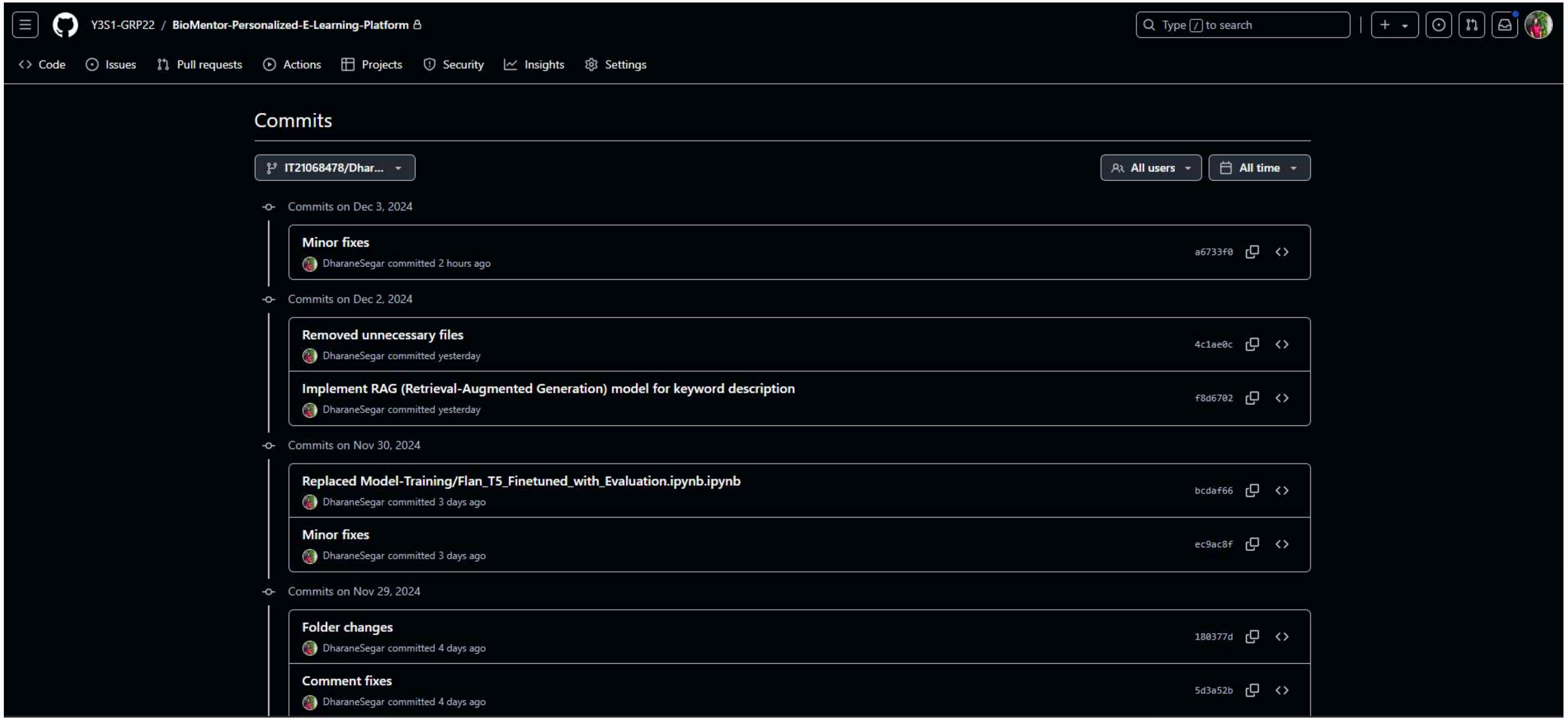
```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

# Path to your fine-tuned model
model_path = 'D:/Downloads/RP/Summarization/flan_t5_finetuned_model-20241119T102614Z-001/flan_t5_finetuned_model'  # Update with the correct path

# Load the tokenizer and model
tokenizer = AutoTokenizer.from_pretrained(model_path)
model = AutoModelForSeq2SeqLM.from_pretrained(model_path)
```



```
from sentence_transformers import SentenceTransformer
import numpy as np
import faiss

# Load embedding model
embedder = SentenceTransformer('all-MiniLM-L6-v2')

# Generate embeddings for summarization dataset (long texts)
summary_embeddings = embedder.encode(long_texts)

# Generate embeddings for notes dataset
notes_embeddings = embedder.encode(notes_content)

# Combine all content and embeddings for FAISS indexing
all_content = long_texts + notes_content
all_embeddings = np.concatenate([summary_embeddings, notes_embeddings], axis=0)

# Convert embeddings to a float32 NumPy array
all_embeddings_array = np.array(all_embeddings).astype("float32")

# Create and populate the FAISS index
common_index = faiss.IndexFlatL2(all_embeddings_array.shape[1])
common_index.add(all_embeddings_array)
```

```
def generate_summary_for_long_text(long_text, min_words=150, max_words=350):
    from textwrap import wrap

    # Helper function to chunk text
    def chunk_text(text, max_tokens=500):
        words = text.split()
        chunks = [' '.join(words[i:i + max_tokens]) for i in range(0, len(words), max_tokens)]
        return chunks

    # Check if input exceeds the max token limit
    max_input_words = 390  # ~512 tokens
    if len(long_text.split()) > max_input_words:
        # Chunk the input into smaller parts
        chunks = chunk_text(long_text, max_tokens=max_input_words)

        # Generate a summary for each chunk and combine the results
        summaries = [generate_summary_for_long_text(chunk, min_words, max_words) for chunk in chunks]
        combined_summary = " ".join(summaries)

        # Ensure the combined summary fits within the final word range
        return truncate_to_word_count(combined_summary, max_words)

    # For shorter inputs, generate the summary directly
    prompt = (
        f"Generate a concise, well-structured, and grammatically correct summary for the following content:\n\n"
        f"{long_text}:\n\nSummary:"
```

GitHub Commits



SLIIT FACULTY OF COMPUTING

Tasks to be done

- Text Extraction from Various Types of Documents
- Customizable Word Count Feature
- Voice Output Integration
- Backend Implementation in Two Architectures
- Frontend Development
- API Development

References

[1] Abstractive Text Summarization Using BART. [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 15 May 2024]

[2] NLP-Enhanced Long Document Summarization: A Comprehensive Approach for Information Condensation. [ONLINE] https://ieeexplore.ieee.org/document/10551101 [Accessed 20 May 2024]

[3] Topic level summary generation using BERTinduced Abstractive Summarization Model. [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 01 June 2024]

[4] Speech-to-Text and Text-to-Speech Recognition using Deep Learning. [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 05 June 2024]

[5] A Novel Text To Speech Conversion Using Hierarchical Neural Network. [ONLINE] https://ieeexplore.ieee.org/document/10533516 [Accessed 05 June 2024]

[6] An Abstractive Summarization and Conversation Bot using T5 and its Variants. [ONLINE]. https://ieeexplore.ieee.org/document/9972639 [Accessed 01 Aug 2024]

IT21264634

Sujitha.S
Software Engineering

LLM BASED ADAPTIVE QUIZ PLATFORM TO IMPROVE MCQ ANSWERING SKILLS IN BIOLOGY FOR STUDENTS

Introduction

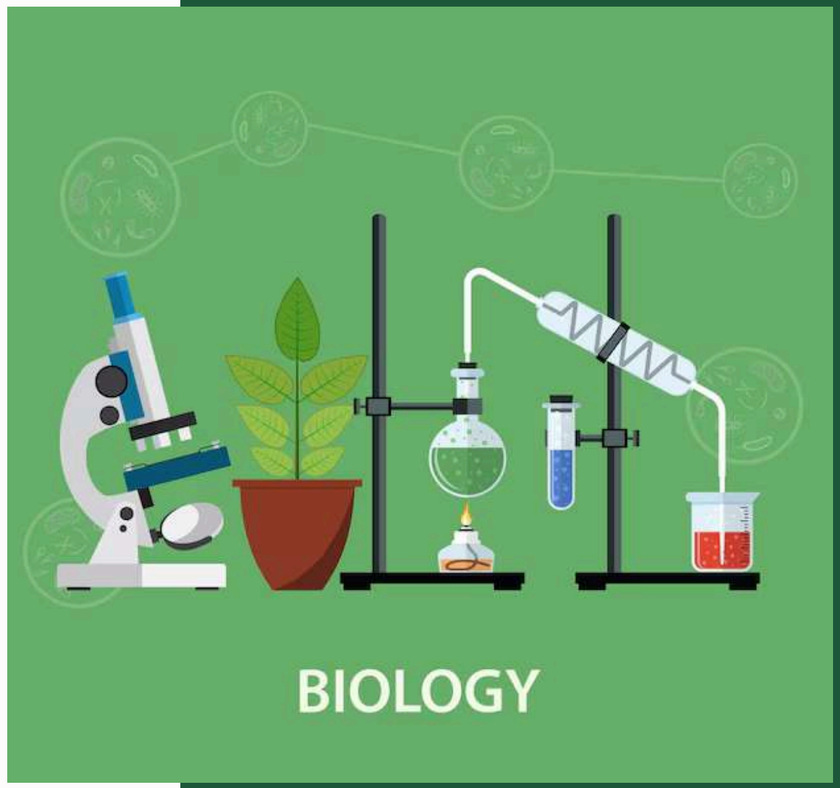
01 Background

02 Research Question

03 Research Gap

04 Main and Sub Objectives

05 Methodology



BACKGROUND

Increasing demand for personalized learning experiences in education.

Need for adaptive assessment tools that cater to individual learning paces and capabilities.

Technological advancements enabling real-time data analysis and dynamic content adjustment.

SLIIT
FACULTY OF COMPUTING

IT21264634 |Sujitha. S | 24-25J-257

RESEARCH PROBLEM

01 Traditional quiz platforms fail to provide real-time adaptability based on individual student performance.

02 There is a need for a system that can offer varied levels of difficulty and content tailored to each learner.

03 Ensuring comprehensive assessment and detailed performance tracking is challenging in existing systems.

Research Gap

	Adaptive MCQ Generation	Personalized	Specialized for Sri Lankan a/l syllabus	MCQ Generation
TOOLSADAY	✓	✗	✗	✓
Eklavvya.AI	✗	✗	✗	✓
BIOMENTOR	✓	✓	✓	✓

OBJECTIVES

Objective 1

Create a variety of MCQs categorized by difficulty to stimulate different cognitive skills.

Objective 2

Enhance the platform's ability to identify knowledge gaps and provide targeted practice recommendations.

Main Objective

Develop an intelligent adaptive quiz system that personalizes MCQs based on student performance, integrates government-approved educational resources, and dynamically adjusts difficulty to cater to varying proficiency levels.

Objective 3

Design a system for continuous performance tracking and detailed analysis report generation.

Objective 4

Ensure user-friendly interface and accessibility for diverse learner groups.

IT21264634 |Sujitha. S | 24-25J-257

Methodology

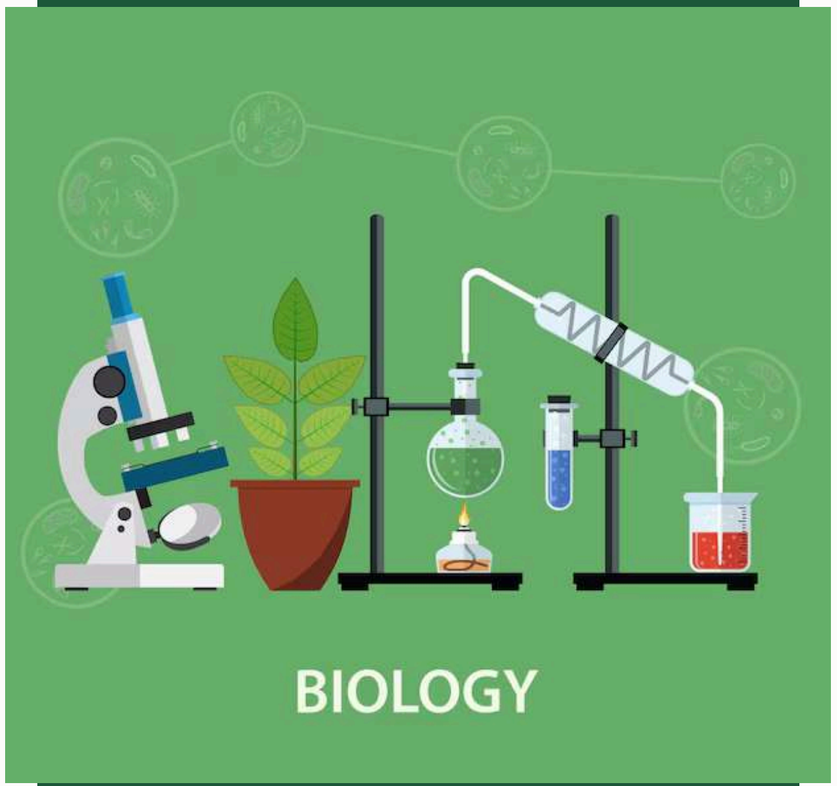
01 System Diagram

02 Tools and Technologies

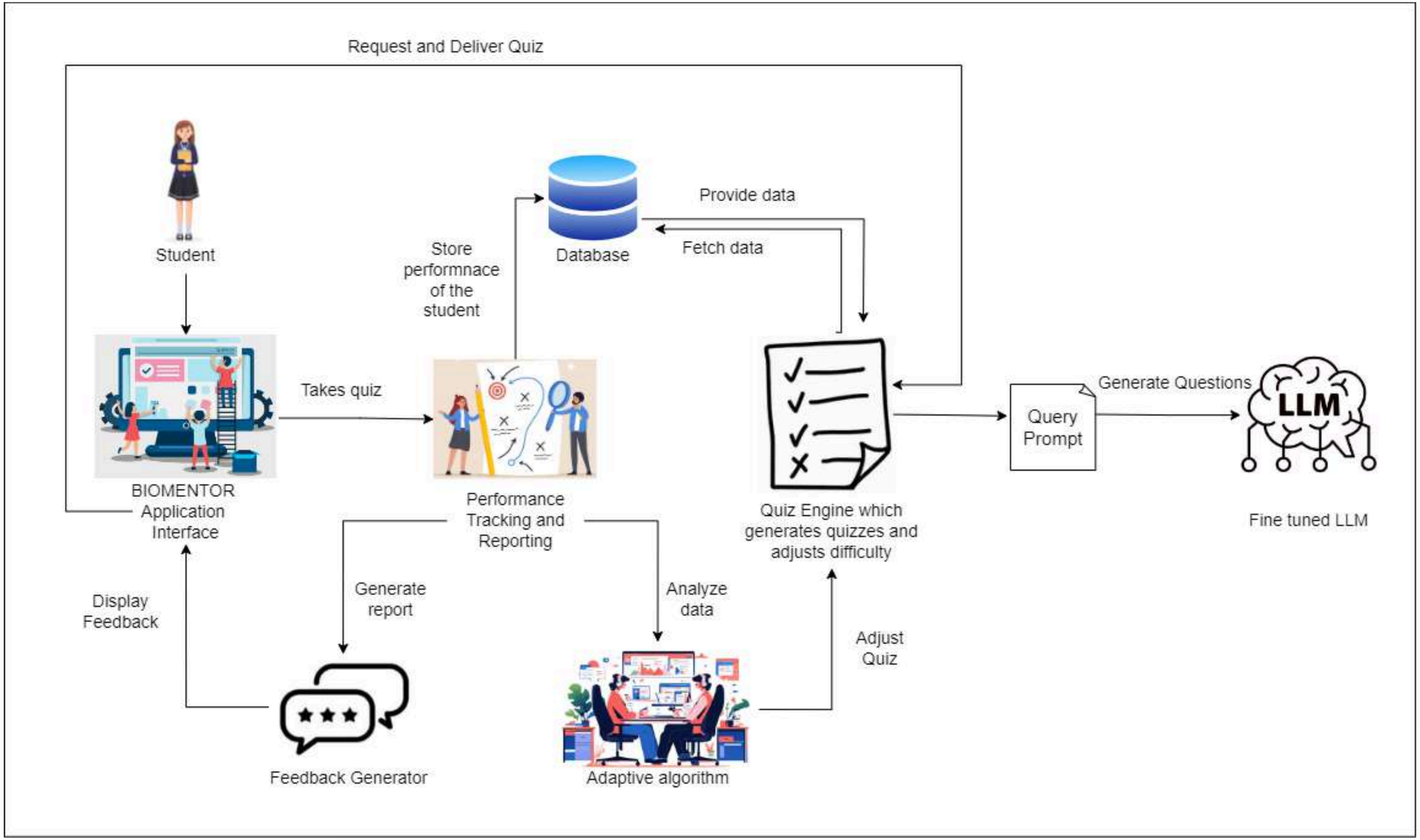
03 Requirements

04 Work Breakdown Structure

05 Gantt Chart



System Diagram



Tools & Technologies

Programming Languages

Python
React Js

Database

Faiss
PostgreSQL
Redis

Project Management

Jira

Other tools

Git
Draw.io
Postman
Figma

Frameworks

Transformers
Pandas
PyTorch
Flask
peft
NumPy
NLTK

SLIIT
FACULTY OF COMPUTING

IT21264634 |Sujitha. S | 24-25J-257

Requirements

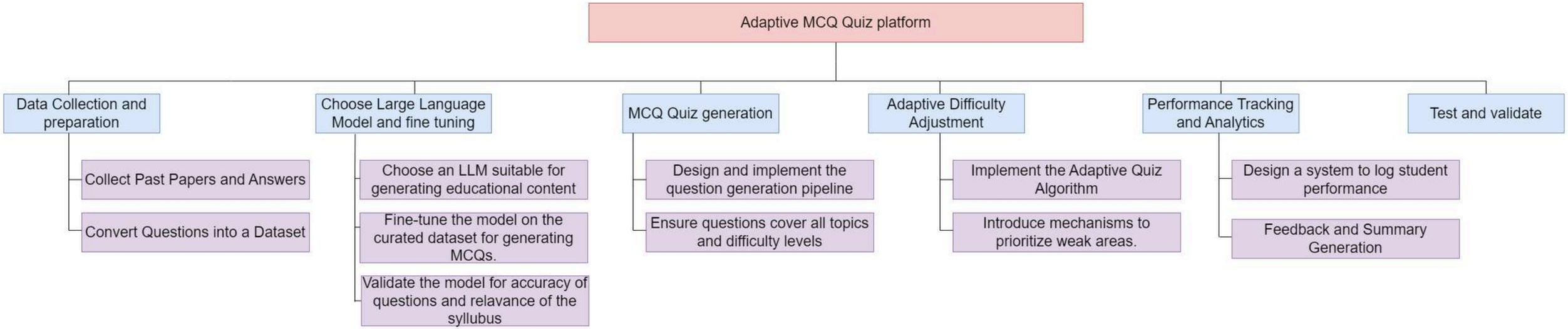
Functional

1. Collect Questions and data from Advanced Level Biology Past Papers
2. Prepare dataset with collected Government Biology Resources for Advanced Level
3. Implement real-time adjustment of question difficulty based on student performance.
4. Track and store individual performance data over time to analyze trends and areas for improvement.
5. Generate detailed performance analysis reports with metrics on accuracy, time taken, and progress.

Non-Functional

1. Compliance
2. Maintainability
3. Usability
4. Scalability
5. Performance

Work Breakdown Structure



Completion of the project

SLIIT
FACULTY OF COMPUTING

IT21264634 |Sujitha. S | 24-25J-257

Data Collection

AutoSave Off merged_mcq_dataset Search

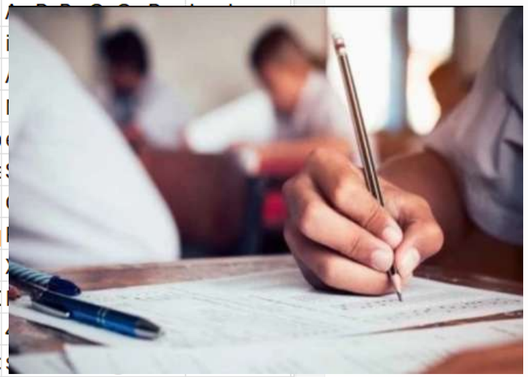
File Home Insert Page Layout Formulas Data Review View Automate Help Acrobat Comments Share

A12 Which of the following statements regarding vascular tissues of plants is correct?

	A	B	C	D	E	F	G	H
1	Question Text	Option 1	Option 2	Option 3	Option 4	Option 5	Correct Answer	Difficulty L
2	A feature common to lysosomes and peroxisomes is?	They are single membr	They transport resid	They contain oxidisir	They are important 1	They digest worn ou	They are single me	easy
3	Two characteristics that can be seen only in living organisms are?	adaptation and growth	movement and irrita	change with time an	metabolism and here	synthesis and deco	metabolism and he	easy
4	Which of the following statements is correct regarding transmission electro	Specimens are magnif	Less electrons may	Living specimens ca	Three-dimensional a	Specimens scatter	Living specimens c	hard
5	Which response which correctly indicates the event and phase in the eukar	DNA replication - G0 p	Synthesis of protein	Chromatin formatior	Production of cellula	Duplication of cent	Synthesis of protei	medium
6	In allosteric regulation of enzymes, which of the following statements is cor	regulatory molecules b	regulatory molecule	an activator molecul	inhibitory molecules	ATP functions as ar	regulatory molecul	medium
7	In ethyl alcohol fermentation, which of the following statements is correct?	one molecule of gluco	pyruvate is reduced	one molecule of CO2	final hydrogen accep	two molecules of A	two molecules of A	easy
8	Which of the following statements regarding glycosis of one molecule of glu	There is a net yield of f	Two hydrogen ions a	It partially depends c	Two NADH molecule	Part of glycolysis ta	Two NADH molecu	medium
9	Some events that took place during the evolution of organisms are as follow	A, B, C and D.	C, A, B and D.	C, B, A and D.	D, A, B and C.	D, B, A and C.	D, B, A and C.	hard
10	Which of the following pairs of organisms have the highest number of comm	Bat and crow.	Lizard and turtle.	Ichthyophis and Tae	Ulva and Pogonatum	Pinus and Cycas.	Lizard and turtle.	easy
11	Which of the following are unique characteristics of some phyla of the kingd	A and C only.	A and D only.	B and C only.	B and D only.	C and D only.	B and C only.	medium
12	Which of the following statements regarding vascular tissues of plants is co	Xylem tissues of ptero	Xylem vessel elemer	Tracheids provide su	Companion cells are	Pits are present be	Xylem tissues of pt	easy
13	Some structures of plants and their functions are shown below: A - Lenticels	A – P, B – R, C – Q.	A - R, B – P, C - P.	A - P, B - Q, C - R.	A - Q, B - P, C - P.	A – R, B-Q, C- R.		
14	Transport of water molecules due to physical adsorption by hydrophilic mat	imbibition.	osmosis.	facilitated diffusion.	bulk flow.	mass flow.		
15	Some steps in the process of opening and closing of stomata are given belo	A, B, C, D, E and F.	A, C, B, D, E and F.	A, C, D, B, E and F.	A, E, B, D, C and F.	A, E, C, D, B and F.		
16	A macronutrient and respectively a micronutrient which cause chlorosis in	Mg and Mn.	Fe and Ni.	P and Mo.	N and S.	Cu and B.		
17	Two plant hormones that promote root formation are?	auxin and gibberellins.	cytokinins and absci	ethylene and auxin.	ethylene and gibbere	cytokinins and gibb		
18	Which of the following statements regarding epithelia is correct?	Stratified squamous e	Pseudostratified col	Simple columnar epi	Simple cuboidal epitl	Simple squamous e		
19	Which of these combinations of the three types of symbiosis seen among or	A only.	B only.	C only.	A and B only.	A and C only		
20	Which is the correct route of blood through the human heart from systemic	Left atrium, bicuspid v	Right atrium. tricusp	Left atrium. tricuspi	Left ventricle. bicusp	Right atrium. bicus		
21	Select the pair/pairs where an increase in (i) causes an increase in (ii. X - (i	X only.	Y only.	Z only.	X and Y only.	X and Z only.		
22	Which of the following indicates the forms that transport the lowest and high	Dissolved CO? - Carba	HCO?? - Carbamino	Carbaminohemoglol	HCO?? - Dissolved C	Dissolved CO? - HC		
23	If the tidal volume. residual volume. inspiratory reserve volume and expirato	1600 ml.	1700 ml.	3600 ml.	4700 ml.	5200 ml.		
24	Which of the following is part of the parasympathetic division of the autonon	inhibits saliva secretio	dilates the pupil of e	relaxes bronchi in lu	stimulates the releas	stimulates gall blac		
25	Which of the following statements regarding human vision is correct?	Changing refractory pc	Convergence occurs	Accommodation is ir	Photopsin in rods pro	Correct perception	Accommodation is	hard

merged_mcq_dataset

Ready Accessibility: Unavailable 100%



SLIIT FACULTY OF COMPUTING

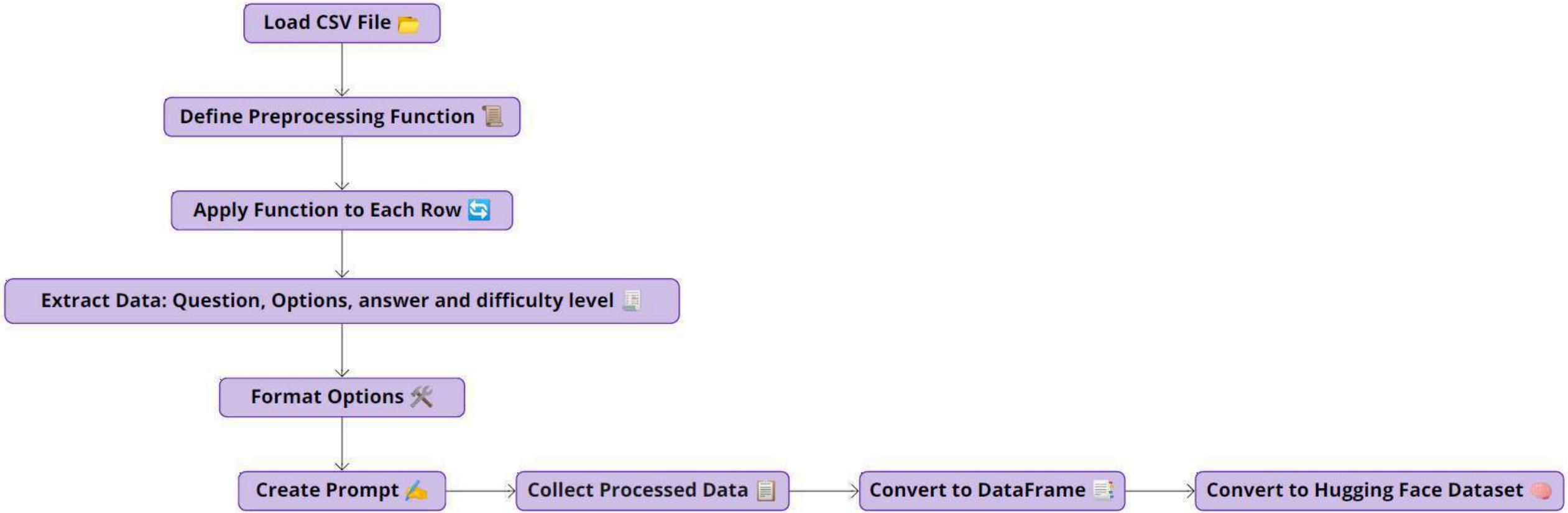
IT21264634 |Sujitha. S | 24-25J-257

Model Selection

Why llama-2-7b-chat-hf?

- Pretrained on Diverse Data
- Fine-Tuning Capability
- Natural Language Understanding
- Customization for Domain-Specific Terminology
- Cost-Effective and Accessible
- Scalability and Efficiency
- Contextual Responses

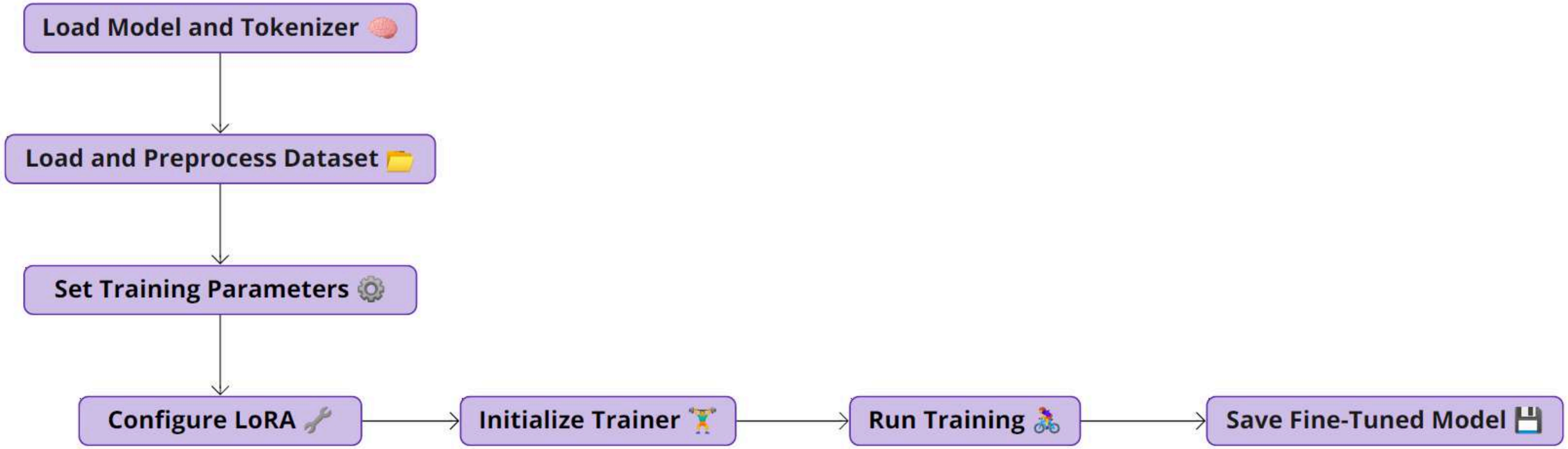
Data pre-processing



Data pre-processing

```python
# Load and preprocess your CSV dataset
df = pd.read_csv(file_path, encoding='ISO-8859-1')

# Define a preprocessing function to format the questions, options, and difficulty
def preprocess_data(row):
    question = row["Question Text"]
    options = [row[f"Option {i}"] for i in range(1, 6) if pd.notna(row[f"Option {i}"])]
    correct_answer = row["Correct Answer"]
    difficulty = row["Difficulty Level"]

    # Format question prompt for training
    options_text = " ".join([f"{chr(65+i)}. {opt.strip()}" for i, opt in enumerate(options)])
    prompt = f"Question ({difficulty}): {question}\nOptions: {options_text}\nChoose the correct answer:"

    return {
        "text": prompt,
        "label": correct_answer
    }

# Apply preprocessing to the dataset and convert to DataFrame
processed_data = df.apply(preprocess_data, axis=1).tolist()
processed_df = pd.DataFrame(processed_data)

# Convert the preprocessed DataFrame into a Hugging Face Dataset format
dataset = Dataset.from_pandas(processed_df)
```

Fine-tuning the LLM

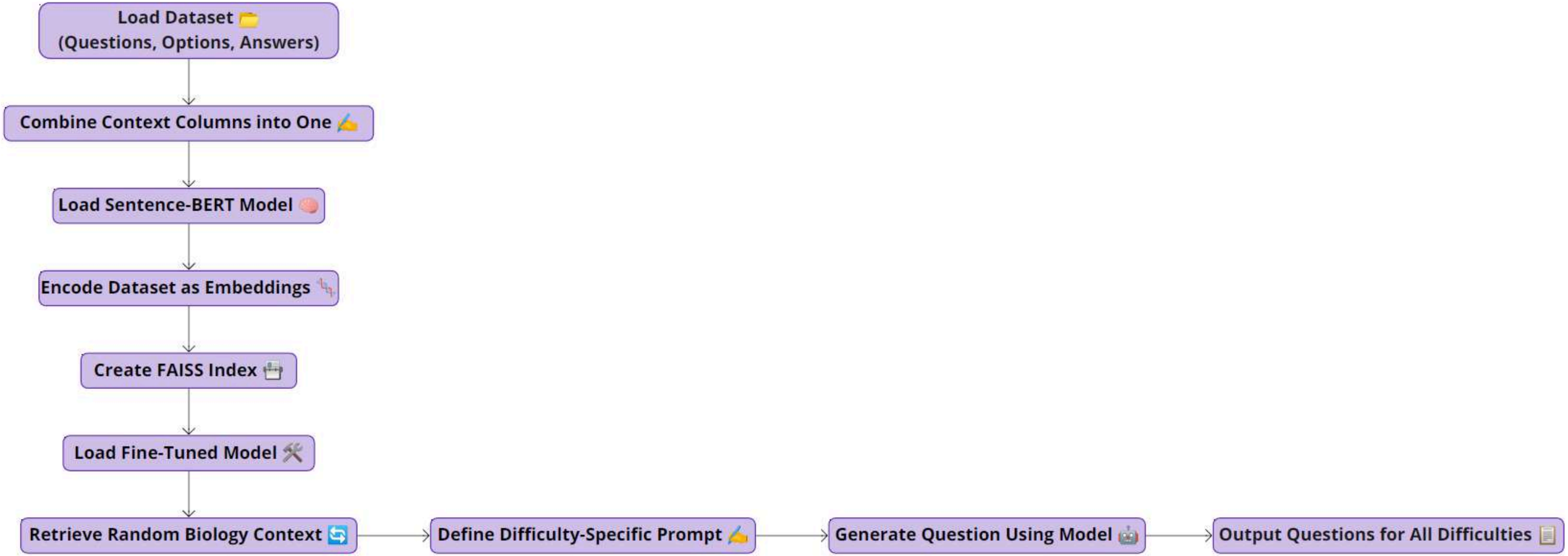


Fine-tuning the LLM

```python
# Configure bitsandbytes
compute_dtype = getattr(torch, bnb_4bit_compute_dtype)
bnb_config = BitsAndBytesConfig(
    load_in_4bit=use_4bit,
    bnb_4bit_quant_type=bnb_4bit_quant_type,
    bnb_4bit_compute_dtype=compute_dtype,
    bnb_4bit_use_double_quant=use_nested_quant,
)

# GPU compatibility check for bfloat16
if compute_dtype == torch.float16 and use_4bit:
    major, _ = torch.cuda.get_device_capability()
    if major >= 8:
        print("=" * 80)
        print("Your GPU supports bfloat16: accelerate training with bf16=True")
        print("=" * 80)

# Load model and tokenizer
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    quantization_config=bnb_config,
    device_map=device_map
)
model.config.use_cache = False
model.config.pretraining_tp = 1

# Load tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
tokenizer.pad_token = tokenizer.eos_token
tokenizer.padding_side = "right"

# LoRA configuration
peft_config = LoraConfig(
    lora_alpha=lora_alpha,
    lora_dropout=lora_dropout,
    r=lora_r,
    bias="none",
    task_type="CAUSAL_LM",
)
```

```python
# Set training parameters
training_arguments = TrainingArguments(
    output_dir=output_dir,
    num_train_epochs=num_train_epochs,
    per_device_train_batch_size=per_device_train_batch_size,
    gradient_accumulation_steps=gradient_accumulation_steps,
    optim=optim,
    save_steps=save_steps,
    logging_steps=logging_steps,
    learning_rate=learning_rate,
    weight_decay=weight_decay,
    fp16=fp16,
    bf16=bf16,
    max_grad_norm=max_grad_norm,
    max_steps=max_steps,
    warmup_ratio=warmup_ratio,
    group_by_length=group_by_length,
    lr_scheduler_type=lr_scheduler_type,
    report_to="tensorboard"
)

# Initialize the trainer with the processed dataset and LoRA configuration
trainer = SFTTrainer(
    model=model,
    train_dataset=dataset,
    peft_config=peft_config,
    dataset_text_field="text",
    max_seq_length=None,  # Or specify max sequence length if needed
    tokenizer=tokenizer,
    args=training_arguments,
    packing=False,
)

# Train the model
trainer.train()
```

RAG Implementation



RAG Implementation

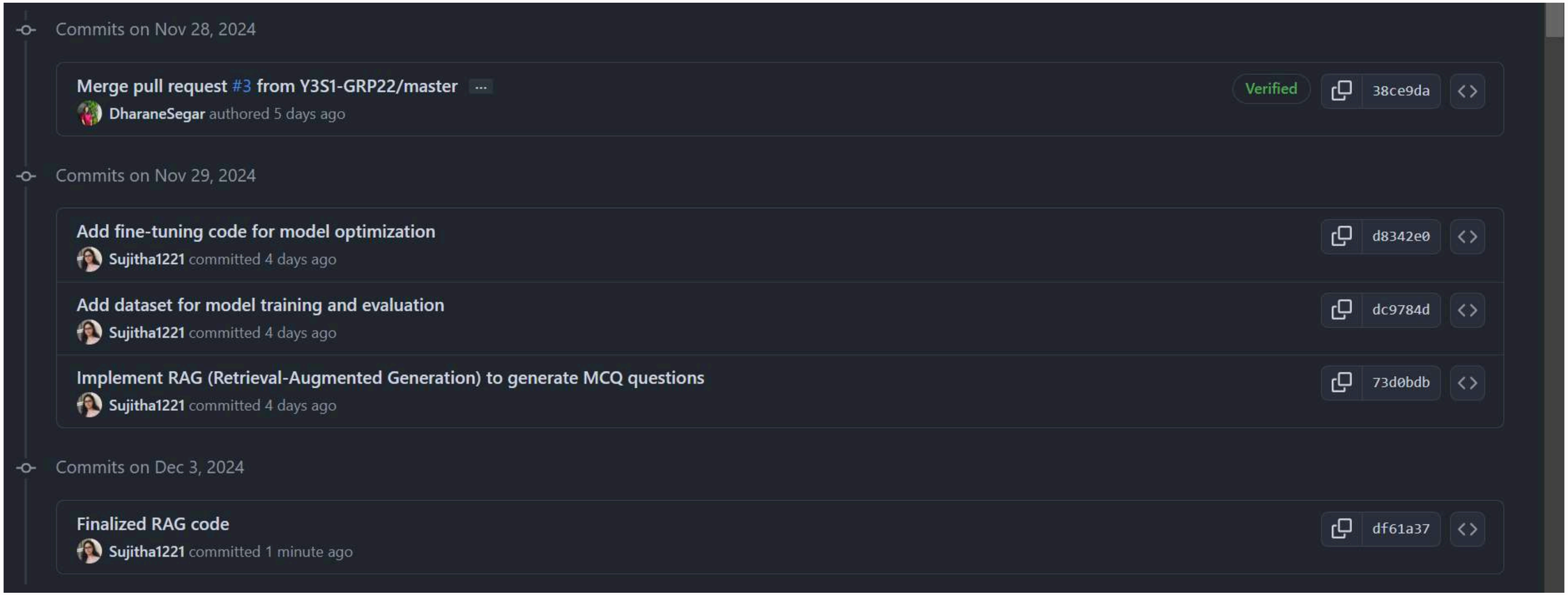
```python
import pandas as pd
import numpy as np
import faiss
from transformers import AutoModelForCausalLM, AutoTokenizer
from sentence_transformers import SentenceTransformer

# Load the dataset
file_path = "/content/drive/My Drive/MCQ Question Generation/merged_mcq_dataset.csv"
dataset = pd.read_csv(file_path, encoding="latin1")
dataset.fillna("", inplace=True)

# Combine all context columns into one
dataset['Combined'] = (
    dataset['Question Text'].astype(str) + " " +
    dataset['Option 1'].astype(str) + " " +
    dataset['Option 2'].astype(str) + " " +
    dataset['Option 3'].astype(str) + " " +
    dataset['Option 4'].astype(str) + " " +
    dataset['Option 5'].astype(str)
)

# Load Sentence-BERT for semantic search
semantic_model = SentenceTransformer('all-MiniLM-L6-v2')
embeddings = semantic_model.encode(dataset['Combined'].tolist(), show_progress_bar=True)

# Create a FAISS index
dimension = embeddings.shape[1]
faiss_index = faiss.IndexFlatL2(dimension)
faiss_index.add(np.array(embeddings))

# Load the fine-tuned model
model_name = "/content/drive/My Drive/MCQ Question Generation/saved_model"  # Update this path
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
generation_model = AutoModelForCausalLM.from_pretrained(model_name)

# Retrieve diverse biology contexts using FAISS
def retrieve_diverse_contexts_faiss(k=3):
    # Randomly select embeddings from the FAISS index
    random_indices = np.random.choice(len(dataset), k, replace=False)
    random_embeddings = np.array([embeddings[i] for i in random_indices])

    # Perform a FAISS search on these random embeddings
    _, indices = faiss_index.search(random_embeddings, k=1)  # Retrieve nearest neighbor for each
    return dataset.iloc[indices.flatten()]['Combined'].tolist()
```

```python
# Define the function to generate questions for all difficulties
def generate_biology_questions():
    difficulties = ["easy", "medium", "hard"]
    questions = {}

    for difficulty in difficulties:
        # Retrieve a random biology context
        retrieved_context = retrieve_context_biology(k=1)
        context = " ".join(retrieved_context)

        # Create a biology-specific prompt
        # Create a refined biology-specific prompt with difficulty-specific instructions
        prompt = f"""
        Based on the following biology context:
        {context}

        Your task is to generate a **biology multiple-choice question** for the **{difficulty} level**:

        - **For Easy Level**: The question should test basic biology concepts, simple definitions, or fundamental facts that are straightforward and easy to recall.
        - **For Medium Level**: The question should involve intermediate biology concepts, processes, or applications that require some reasoning or understanding of relationships between concepts.
        - **For Hard Level**: The question should test advanced biology concepts, detailed mechanisms, or require critical thinking and analysis of biological principles.

        - Provide exactly **five distinct answer options** labeled a, b, c, d, and e.
        - Only one answer option should be correct.
        - Clearly indicate the correct answer.

        Please output in the following format only:
        - Question: <Your question>
        - a) <Option 1>
        - b) <Option 2>
        - c) <Option 3>
        - d) <Option 4>
        - e) <Option 5>
        - Correct Answer: <Correct option letter>

        Do not include any additional text, explanations, or examples.
        """

        # Tokenize and generate
        input_ids = tokenizer.encode(prompt, return_tensors="pt", padding=True, truncation=True)
        output = generation_model.generate(input_ids, max_length=500, temperature=0.7, top_k=50, top_p=0.85)

        # Decode the response
        question = tokenizer.decode(output[0], skip_special_tokens=True)
        questions[difficulty] = question

    return questions

# Example usage
questions = generate_biology_questions()

# Print the questions
for difficulty, question in questions.items():
    print(f"Difficulty: {difficulty.capitalize()}\n{question}\n")
```

SLIIT
FACULTY OF COMPUTING

IT21264634 |Sujitha. S | 24-25J-257

GitHub Commits



Tasks to be done

- Design the Adaptive Quiz Algorithm
- Redefine dynamic Question Creation Workflow
- Create user Performance Tracking System
- Implement Feedback and Summary Generation
- UI/UX Design and front-end development
- Frontend and Backend Integration

References

[1] Automatic Question Answer Generation using T5 and NLP. [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 13 May 2024]

[2] Generation of Multiple-Choice Questions From Textbook Contents of School-Level Subjects. [ONLINE] https://ieeexplore.ieee.org/document/9964056 [Accessed 20 May 2024]

[3] Automatic question generation for intelligent tutoring systems. [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 06 June 2024]

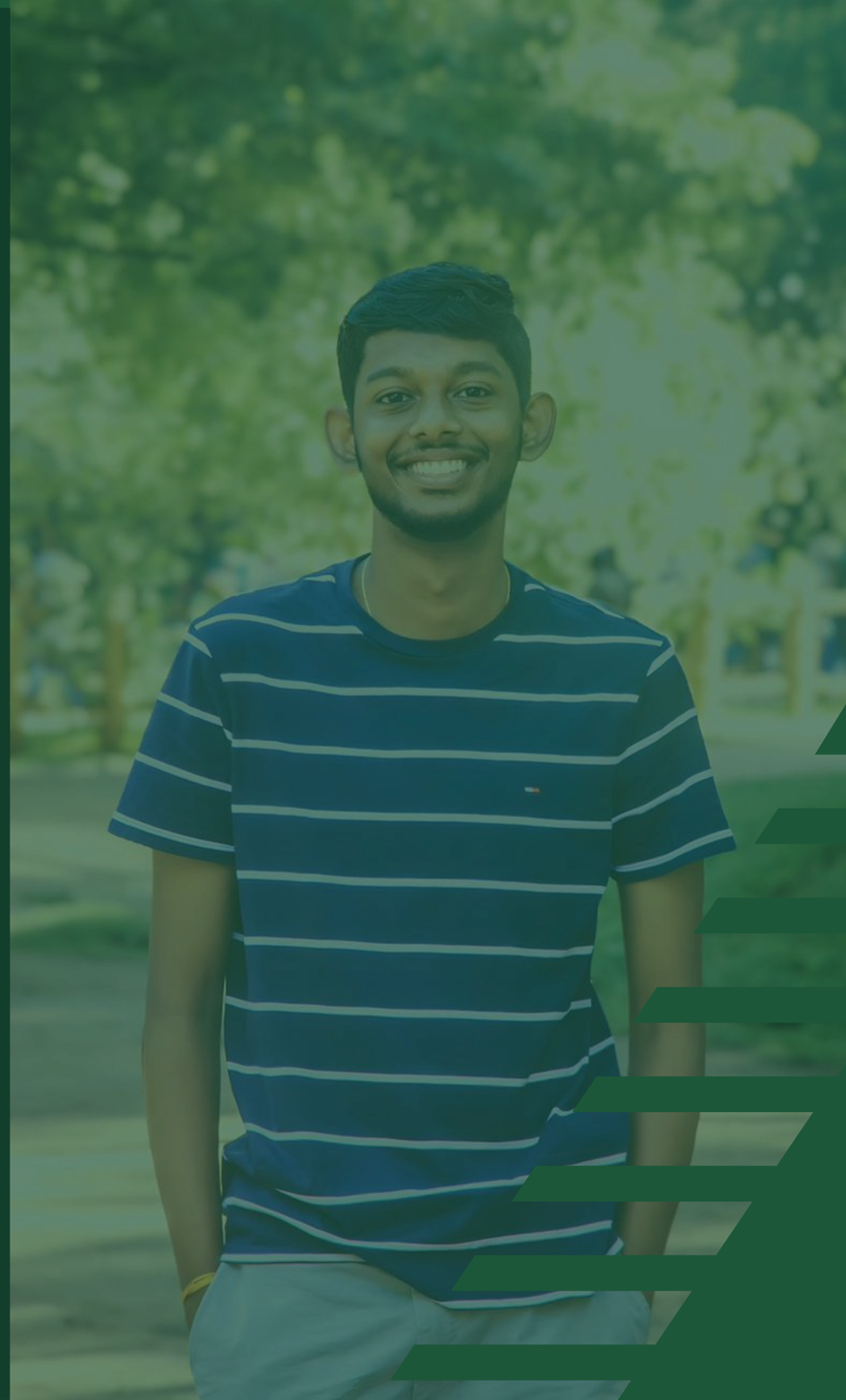
[4] MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning. [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 05 June 2024]

[5] Generation of Multiple Choice Questions from Indian Educational Text. [ONLINE] https://ieeexplore.ieee.org/document/10270551 [Accessed 16 June 2024

IT21204302

Sajeevan S
Software Engineering

LLM BASED PROVIDE ANSWERS FOR STRUCTURED AND ESSAY TYPE OF QUESTIONS AND EVALUATE ANSWERS BASED ON APPROVED RESOURCES.



Introduction

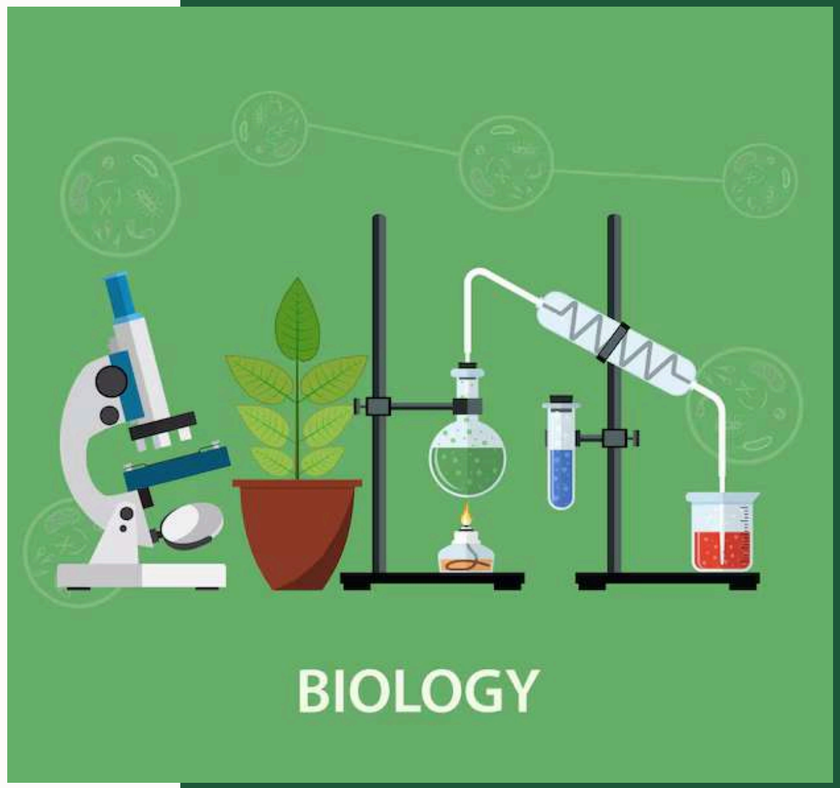
01 Background

02 Research Question

03 Research Gap

04 Main and Sub Objectives

05 Methodology



BACKGROUND

The Importance of Independent Learning in Modern Education

The Challenges of Self-Evaluation in structured essay-type answering

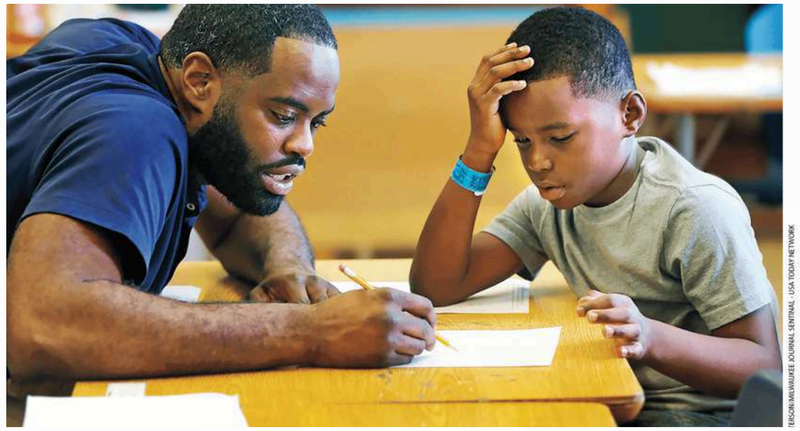
RESEARCH PROBLEM

01

How can technology-based tools be developed and utilized to assist students in improving their essay-writing skills independently, without relying on mentor support?

What strategies can students employ to self-evaluate and improve their structured essay-type answers in the absence of mentor guidance?

02



Research Gap

	For Srilankan A/L Bio syllabus	*Answer based on the Srilankan A/L system*	*Answer Evaluation and suggestion*
p	✗	✗	✓
Claude	✗	✗	✓
BIOMENTOR	✓	✓	✓

OBJECTIVES

Objective 1

Answer Generation for a Question

Objective 2

Enhance Self-Directed Learning in Advanced-Level Biology

Main Objective

If students provide a question, the system will generate an answer, and if students also provide their corresponding answer, the system will evaluate their response and offer suggestions to improve the answer.

Objective 3

Provide Suggestions for Improvement

Objective 4

Provide Feedback for Answers

IT21204302 |Sajeevan. S | 24-25J-257

Methodology

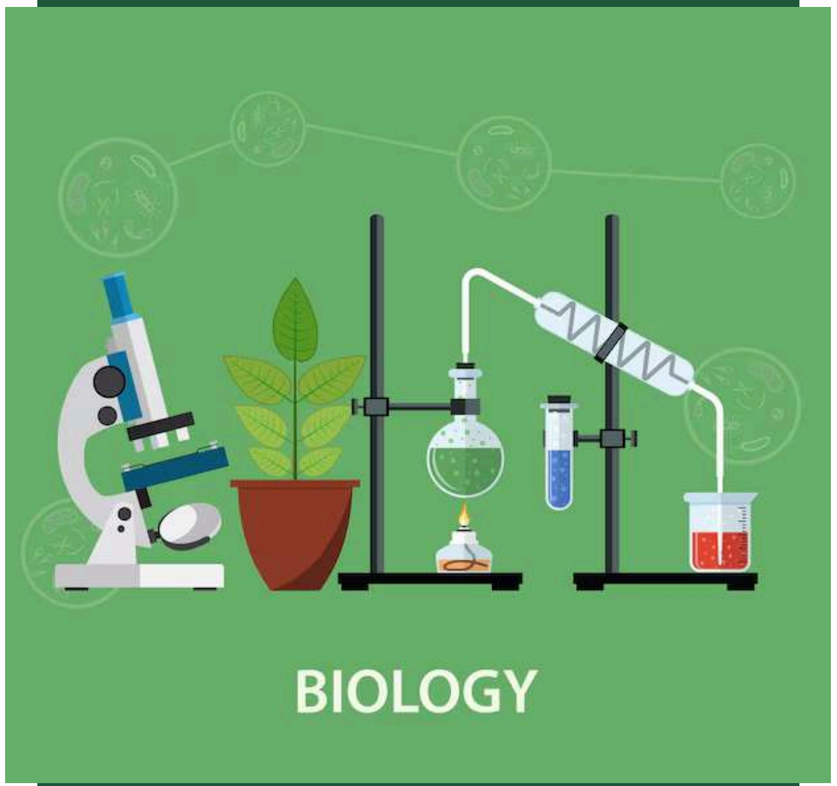
01 System Diagram

02 Tools and Technologies

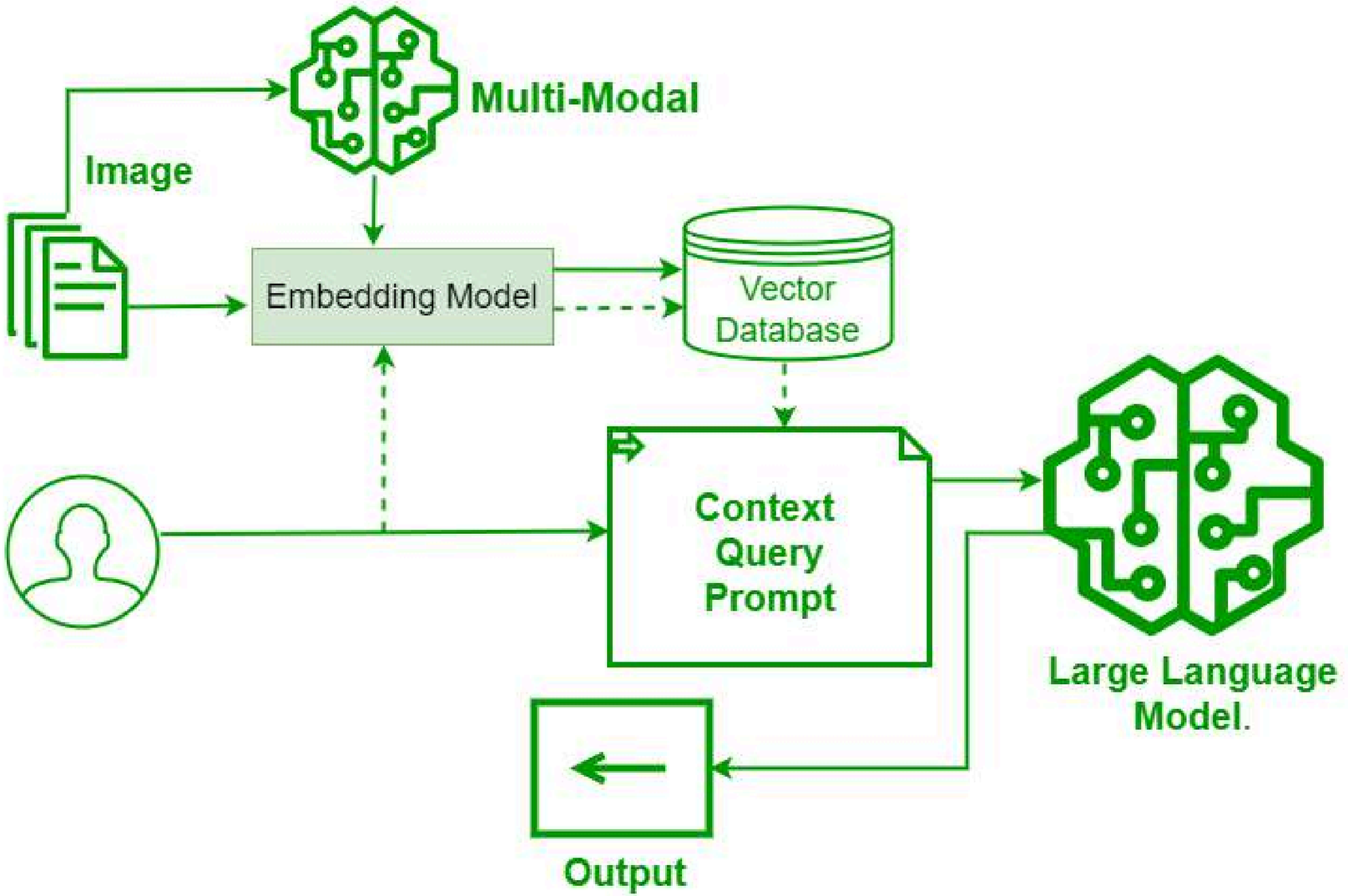
03 Requirements

04 Work Breakdown Structure

05 Gantt Chart



System Diagram



Tools & Technologies

Programming Languages

Python
React JS

Database

Faiss
Mongo DB

Project Management

Jira

Frameworks

Transformer model
Flask
Pytorch
OCR

Other tools

Git
Draw.io
Postman

Requirements

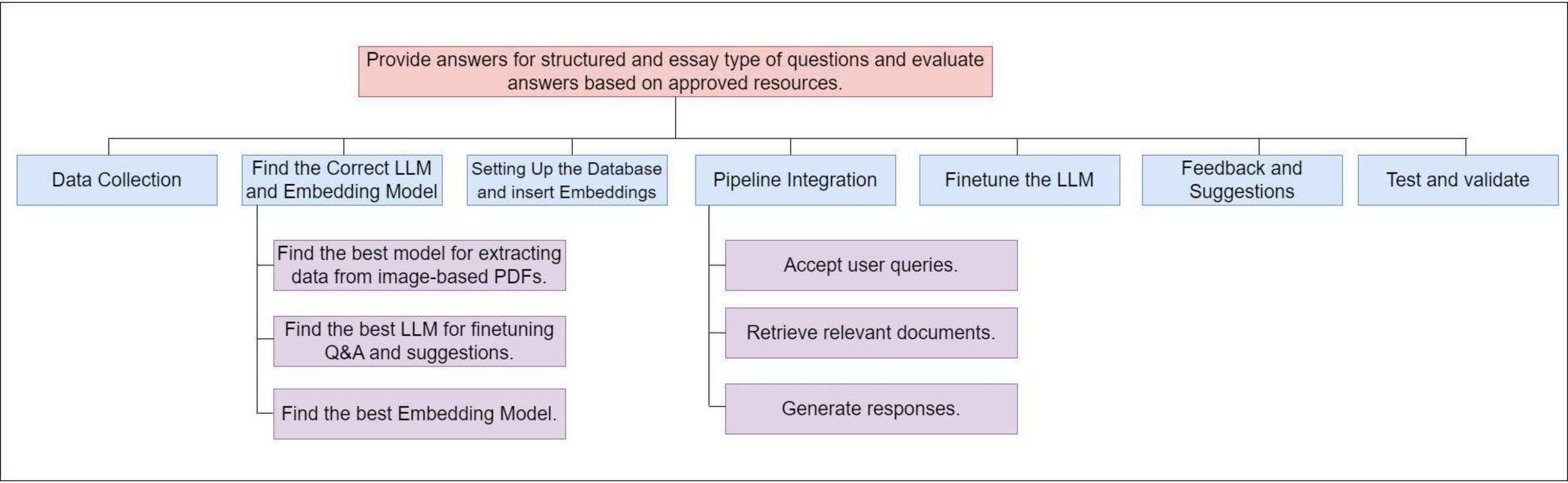
Functional

1. Collect Dataset from Advanced Level Biology Past Papers
2. Update the database with government biology resources and Advanced Level questions and answers.
3. Evaluate Answers and Provide Improvement Suggestions
4. Answer Generation for Structured and Essay Questions

Non-Functional

1. Accuracy
2. Performance
3. Availability
4. Usability

Work Breakdown Structure

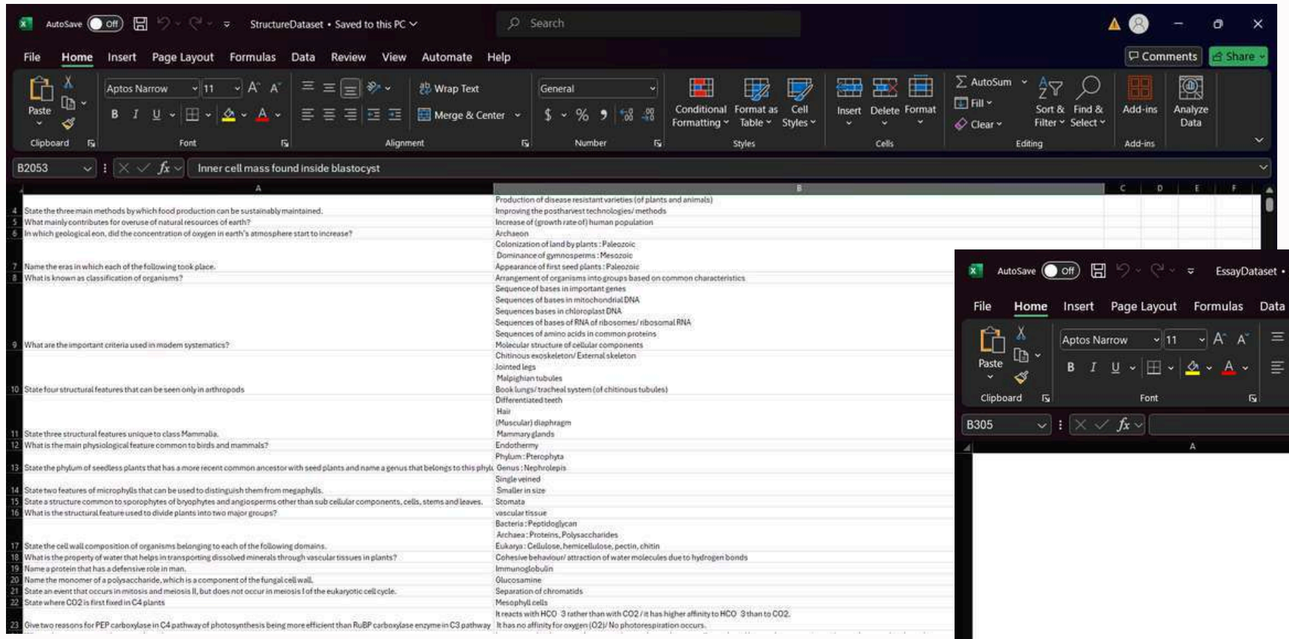


Completion of the project

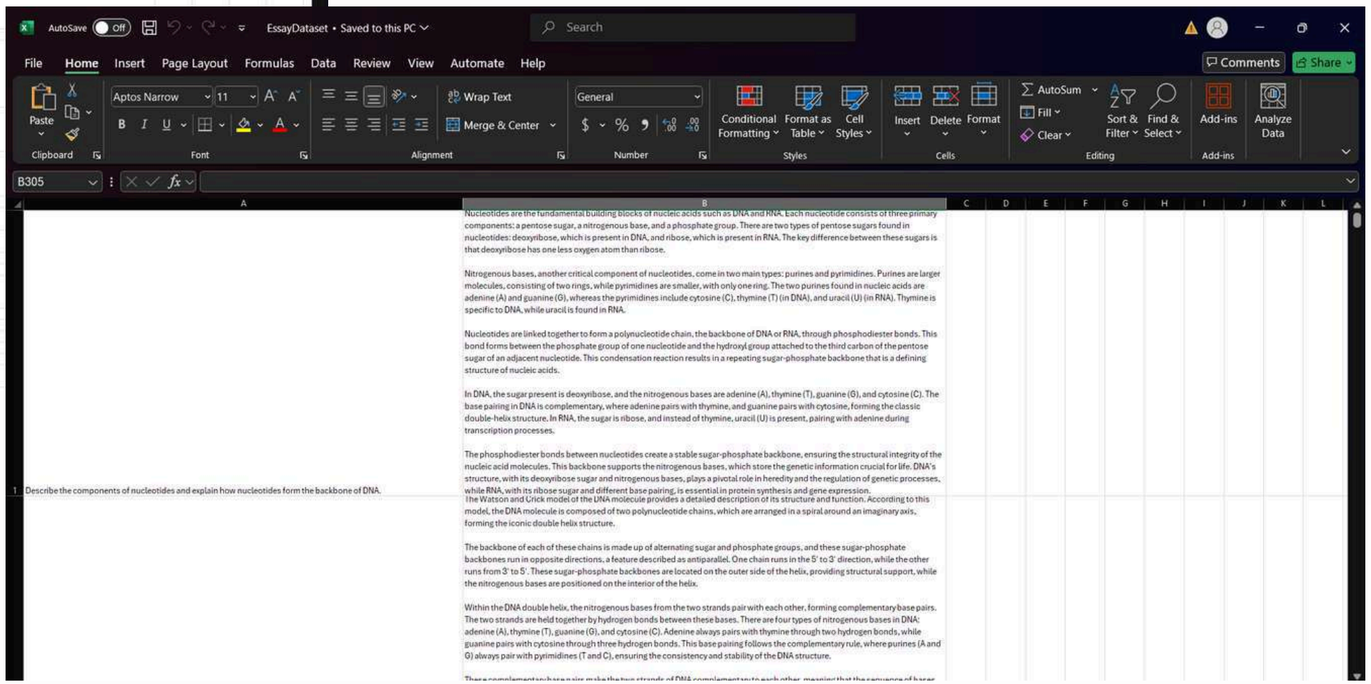
SLIIT
FACULTY OF COMPUTING

IT21204302 |Sajeevan. S | 24-25J-257

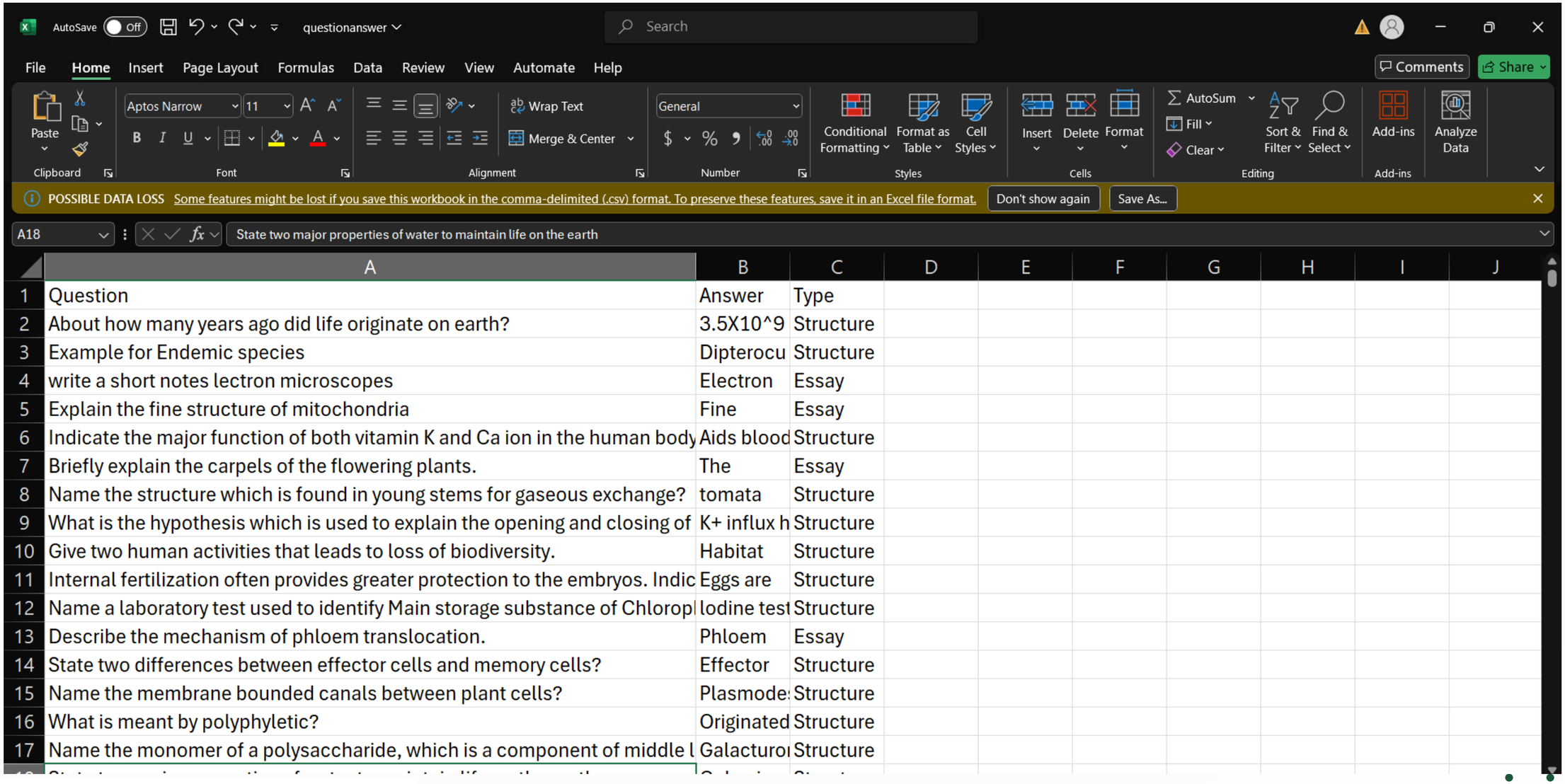
Data Collection



		Production of disease resistant varieties (of plants and animals)
4	State the three main methods by which food production can be sustainably maintained.	Improving the postharvest technologies/ methods
5	What mainly contributes for overuse of natural resources of earth?	Increase of (growth rate of) human population
6	In which geological eon, did the concentration of oxygen in earth's atmosphere start to increase?	Archaeon
		Colonization of land by plants : Paleozoic
		Dominance of gymnosperms : Mesozoic
7	Name the eras in which each of the following took place.	Appearance of first seed plants : Paleozoic
8	What is known as classification of organisms?	Arrangement of organisms into groups based on common characteristics
		Sequence of bases in important genes
		Sequences of bases in mitochondrial DNA
		Sequences bases in chloroplast DNA
		Sequences of bases of RNA of ribosomes/ ribosomal RNA
		Sequences of amino acids in common proteins
9	What are the important criteria used in modern systematics?	Molecular structure of cellular components
		Chitinous exoskeleton/ External skeleton
		Jointed legs
		Malpighian tubules
10	State four structural features that can be seen only in arthropods	Book lungs/ tracheal system (of chitinous tubules)
		Differentiated teeth
		Hair
		(Muscular) diaphragm
11	State three structural features unique to class Mammalia.	Mammary glands
12	What is the main physiological feature common to birds and mammals?	Endothermy
		Phylum : Pterophyta
13	State the phylum of seedless plants that has a more recent common ancestor with seed plants and name a genus that belongs to this phyla	Genus : Nephrolepis
		Single veined
14	State two features of microphylls that can be used to distinguish them from megaphylls.	Smaller in size
15	State a structure common to sporophytes of bryophytes and angiosperms other than sub cellular components, cells, stems and leaves.	Stomata
16	What is the structural feature used to divide plants into two major groups?	vascular tissue
		Bacteria : Peptidoglycan
		Archaea : Proteins, Polysaccharides
17	State the cell wall composition of organisms belonging to each of the following domains.	Eukarya : Cellulose, hemicellulose, pectin, chitin
18	What is the property of water that helps in transporting dissolved minerals through vascular tissues in plants?	Cohesive behaviour/ attraction of water molecules due to hydrogen bonds
19	Name a protein that has a defensive role in man.	Immunoglobulin
20	Name the monomer of a polysaccharide, which is a component of the fungal cell wall.	Glucosamine
21	State an event that occurs in mitosis and meiosis II, but does not occur in meiosis I of the eukaryotic cell cycle.	Separation of chromatids
22	State where CO2 is first fixed in C4 plants	Mesophyll cells
		It reacts with HCO 3 rather than with CO2 / it has higher affinity to HCO 3than to CO2.
23	Give two reasons for PEP carboxylase in C4 pathway of photosynthesis being more efficient than RuBP carboxylase enzyme in C3 pathway	It has no affinity for oxygen (O2)/ No photorespiration occurs.



Data Collection



AutoSave Off | questionanswer | Search

File Home Insert Page Layout Formulas Data Review View Automate Help | Comments | Share

Aptos Narrow | 11 | Wrap Text | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | AutoSum | Fill | Clear | Sort & Filter | Find & Select | Add-ins | Analyze Data

Clipboard | Font | Alignment | Number | Styles | Cells | Editing | Add-ins

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format. Don't show again | Save As...

A18 | State two major properties of water to maintain life on the earth

	A	B	C	D	E	F	G	H	I	J
1	Question	Answer	Type							
2	About how many years ago did life originate on earth?	3.5X10^9	Structure							
3	Example for Endemic species	Dipterocu	Structure							
4	write a short notes lectron microscopes	Electron	Essay							
5	Explain the fine structure of mitochondria	Fine	Essay							
6	Indicate the major function of both vitamin K and Ca ion in the human body	Aids blood	Structure							
7	Briefly explain the carpels of the flowering plants.	The	Essay							
8	Name the structure which is found in young stems for gaseous exchange?	tomata	Structure							
9	What is the hypothesis which is used to explain the opening and closing of	K+ influx h	Structure							
10	Give two human activities that leads to loss of biodiversity.	Habitat	Structure							
11	Internal fertilization often provides greater protection to the embryos. Indic	Eggs are	Structure							
12	Name a laboratory test used to identify Main storage substance of Chlorop	Iodine test	Structure							
13	Describe the mechanism of phloem translocation.	Phloem	Essay							
14	State two differences between effector cells and memory cells?	Effector	Structure							
15	Name the membrane bounded canals between plant cells?	Plasmode	Structure							
16	What is meant by polyphyletic?	Originated	Structure							
17	Name the monomer of a polysaccharide, which is a component of middle l	Galacturo	Structure							

Model Selection

Why Llama 3?

An instruct fine-tuned version of the 8B model that is optimized for specific tasks. For instance, it can be used to create educational tools that explain complex subjects.

Llama 3 has a greater capacity to learn nuanced patterns and adapt to specific tasks compared to other models.

Llama 3 is highly customizable, making it ideal for domain-specific tasks like academic content generation and adaptive learning applications.

Llama 3 generates contextually accurate, meaningful responses while avoiding repetition and irrelevant content.

Model Selection

Issues with Alternatives

1. T5 Small:
   - Lacks clarity in answers due to its limited model size.
2. unsloth_gemma_2b_bnb_4bit:
   - Produces similar responses for different question types, indicating poor adaptability.
   - Lacks diversity in output.
3. unsloth_mistral_7b:
   - Includes irrelevant URLs and citations in essay-type answers.

Fine-tuning the LLM

Environment Setup

- Install libraries (`torch`, `transformers`, `datasets`, etc.).
- Configure device (`cuda` or `cpu`).

Dataset Preparation

- Format dataset into structured text (`Type`, `Question`, `Answer`).
- Convert to Hugging Face `Dataset` and tokenize (max length: 512).
- Split into training (80%) and validation (20%) sets.

Authentication and Model Configuration

- Log into Hugging Face Hub.
- Load `meta-llama/Meta-Llama-3-8B-Instruct` model with 4-bit quantization using `BitsAndBytesConfig`.
- Add `<|pad|>` token to the tokenizer.

LoRA Configuration

- Set up LoRA parameters targeting `q_proj` and `v_proj`.
- Prepare the model for k-bit training.

Fine-Tuning

- Use `SFTTrainer` for training:
- Save the fine-tuned model locally.

Merging LoRA Weights

- Merge LoRA weights with the base model.
- Save the merged model and tokenizer locally.

Fine-tuning the LLM

```python
lora_config = LoraConfig(
    r=16,
    lora_alpha=32,
    lora_dropout=0.1,
    target_modules=["q_proj", "v_proj"],
    bias="none",
    task_type="CAUSAL_LM"
)

# Prepare model for LoRA training
model = prepare_model_for_kbit_training(model)
model = get_peft_model(model, lora_config)

from trl import DataCollatorForCompletionOnlyLM, SFTConfig, SFTTrainer

# Define a response template to pass into DataCollatorForCompletionOnlyLM
response_template = "<|endoftext|>"

# Set up Data Collator with the response template
data_collator = DataCollatorForCompletionOnlyLM(
    tokenizer=tokenizer,
    response_template=response_template,
    return_tensors="pt"
)

# Trainer configuration with minimal logging by setting logging_steps to a high value
sft_config = SFTConfig(
    output_dir="./llama_custom_model",
    max_seq_length=512,
    num_train_epochs=3,
    per_device_train_batch_size=4,
    gradient_accumulation_steps=8,
    save_steps=100,  # Save checkpoints only
    logging_steps=1000000,  # Set to a high number to avoid frequent logging
    evaluation_strategy="no",  # Skip evaluation during training
    learning_rate=1e-4,
```

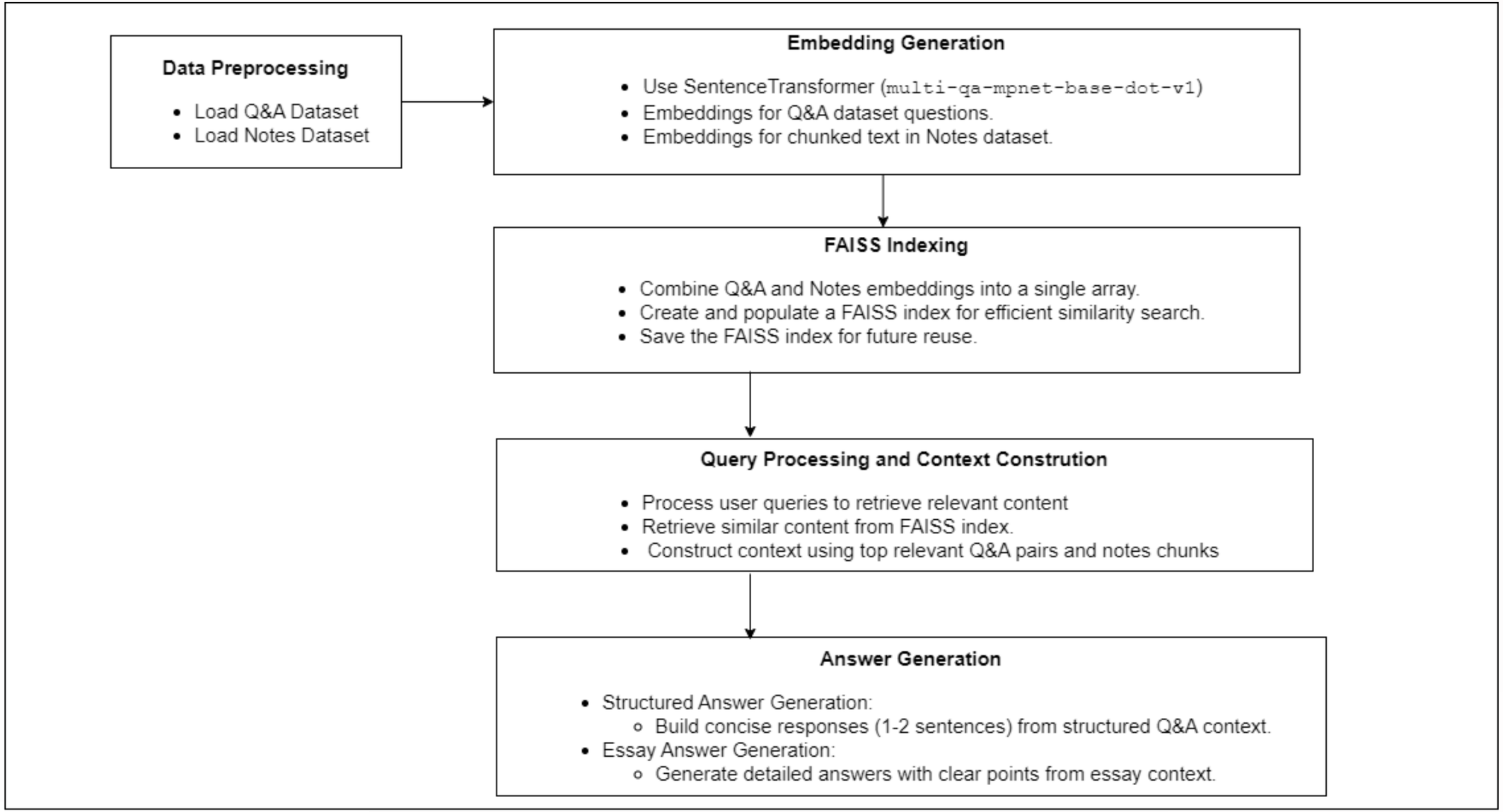
```python
# Trainer configuration with minimal logging by setting logging_steps to a high value
sft_config = SFTConfig(
    output_dir="./llama_custom_model",
    max_seq_length=512,
    num_train_epochs=3,
    per_device_train_batch_size=4,
    gradient_accumulation_steps=8,
    save_steps=100,  # Save checkpoints only
    logging_steps=1000000,  # Set to a high number to avoid frequent logging
    evaluation_strategy="no",  # Skip evaluation during training
    learning_rate=1e-4,
    fp16=True,
    report_to="none"
)

# Initialize Trainer without tracking training loss
trainer = SFTTrainer(
    model=model,
    args=sft_config,
    train_dataset=train_data,
    tokenizer=tokenizer,
    data_collator=data_collator
)

# Start training without loss tracking or logging
trainer.train()
trainer.save_model("./llama_custom_model")
print("Model training completed and saved.")
```

Data pre-processing

```python
# Load the Q&A dataset
qa_df = pd.read_csv('questionanswer.csv', encoding='ISO-8859-1')  # Columns: Question, Answer, Type

# Clean missing values
qa_df['Question'] = qa_df['Question'].fillna('')
qa_df['Answer'] = qa_df['Answer'].fillna('')
qa_df['Type'] = qa_df['Type'].fillna('structured')  # Default to 'structured'

# Remove duplicates
qa_df = qa_df.drop_duplicates()

# Display a preview of the cleaned Q&A dataset
print(qa_df.head())
```

✓ 0.0s

RAG Implementation

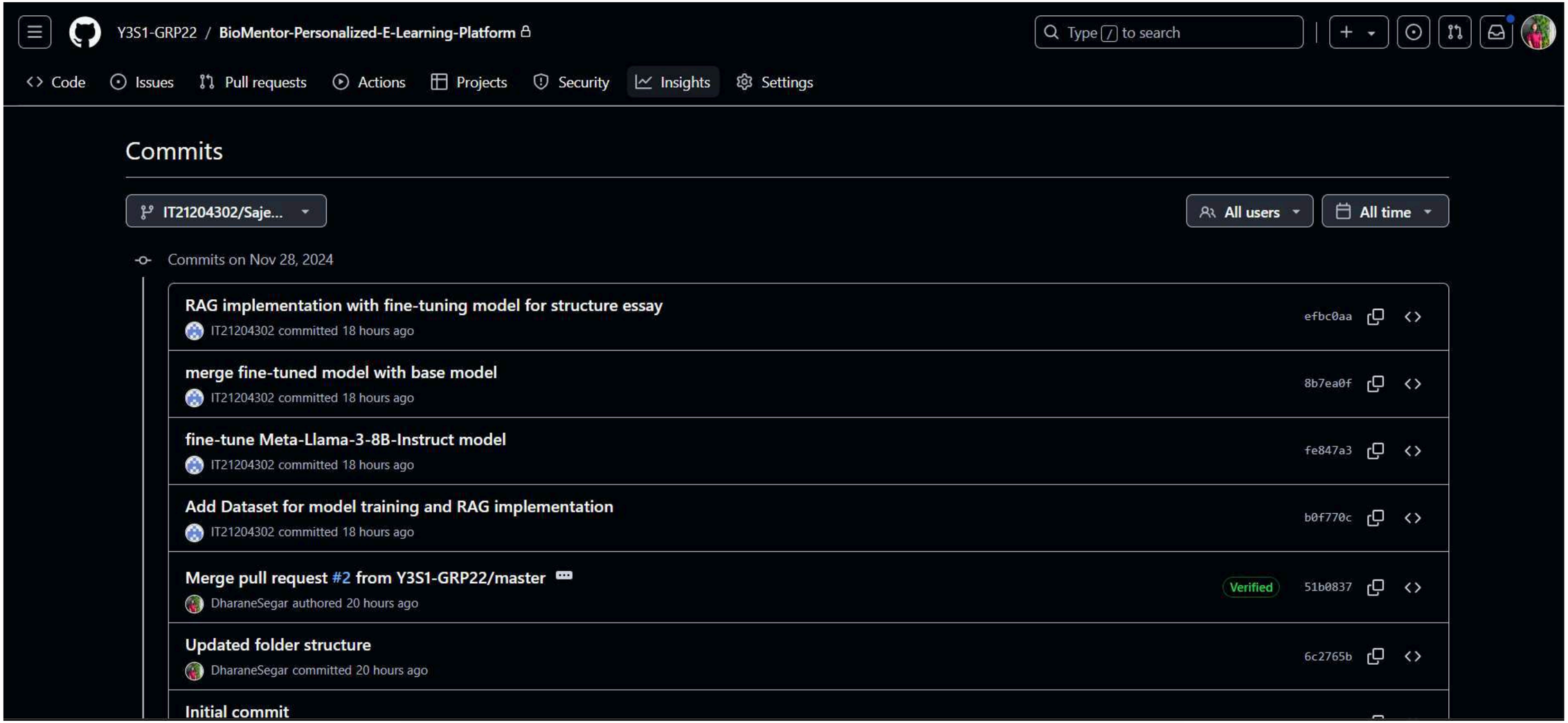


RAG Implementation

```
Tabnine | Edit | Test | Explain | Document | Ask
def generate_structured_answer(query, k=3, max_words=50):
    """
    Generate structured answers with concise and specific responses.
    """
    # Build context for structured questions
    context = construct_context_for_structure(query, k)

    # Prompt for structured questions
    prompt = (
        f"Question: {query}\n\n"
        f"Context:\n{context}\n\n"
        f"Answer the question concisely and accurately in 1-2 sentences.\nAnswer:"
    )

    # Adjust max token length
    input_length = len(generator.tokenizer(prompt)['input_ids'])
    adjusted_max_length = input_length + max_words

    # Generate response
    response = generator(
        prompt,
        max_length=adjusted_max_length,
        truncation=True,
        num_return_sequences=1
    )
    return response[0]["generated_text"]
```

```
Tabnine | Edit | Test | Explain | Document | Ask
def generate_essay_answer(query, k=5, min_words=175, max_words=300):
    """
    Generate essay-style answers with a detailed explanation.
    Ensure the response meets the minimum word count and does not exceed the maximum token count.
    """
    # Construct context with improved filtering
    context = construct_context_for_essay(query, k) or ""

    # Check if context is empty and set the prompt accordingly
    if not context.strip():
        # Fallback prompt when no context is available
        prompt = (
            f"Question: {query}\n\n"
            f"Answer the question in detail, providing a well-reasoned and comprehensive explanation. "
            f"Highlight key features, provide examples, and mention advantages and disadvantages where applicable. "
            f"Ensure your response is grammatically correct and at least {min_words} words long. "
            f"Complete your answer and avoid repetition.\nAnswer:"
        )
    else:
        # Prompt with context
        prompt = (
            f"Question: {query}\n\n"
            f"Context:\n{context}\n\n"
            f"Answer the question using the provided context. "
            f"Focus on relevant details from the context and elaborate as needed. "
            f"Provide clear points and explanations, ensuring your response is at least {min_words} words long and grammatically correct. "
            f"Highlight key features, examples, and advantages/disadvantages where applicable. "
            f"Do not repeat the context verbatim; instead, integrate it meaningfully into the answer. "
            f"Complete your answer and avoid repetition.\nAnswer:"
        )

    # Adjust max token length dynamically
    input_length = len(generator.tokenizer(prompt)['input_ids'])
    estimated_token_count = int(min_words * 0.75)  # Approximate token count for the minimum word count
    adjusted_min_length = input_length + estimated_token_count
```

GitHub Commits



Tasks to be done

- Develop the front-end (UI/UX and client-side logic).
- Implement the back-end (server-side logic and APIs).
- Integrate front-end and back-end for seamless functionality.
- Implement a function to compare user answers and provide evaluations.
- Integrate the function with other components for seamless operation

References

[1] BERT-Based Model for Reading Comprehension Question Answering. [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 12 May 2024]

[2] Question Answering System using NLP and BERT. [ONLINE] https://ieeexplore.ieee.org/document/10551101 [Accessed 22 May 2024]

[3] Question Answering Model Based Conversational Chatbot using BERT Model and Google Dialogflow. . [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 03 June 2024]

[4] BERT-Based Mixed Question Answering Matching Model. [ONLINE] https://ieeexplore.ieee.org/document/9972639 [Accessed 05 June 2024]

[5] Semantic Similarity Detection and Analysis For Text Documents. [ONLINE] https://ieeexplore.ieee.org/document/10533516 [Accessed 15 June 2024

BIOMENTOR

SLIIT
FACULTY OF COMPUTING

BIOMENTOR

THANK YOU

SLIIT
FACULTY OF COMPUTING