Chapter 1

INTRODUCTION

1.1 Introduction

With the rapid integration of technology into education, teaching has become more engaging through tools like projectors, online tutorials, and animated video lessons. However, while instructional methods have evolved, the evaluation of descriptive answer sheets remains largely manual. Faculty members still evaluate exam papers by hand, which is time-consuming, inconsistent, and susceptible to human error. Subjectivity and emotional variation during manual correction can also lead to unfair grading.

To address these challenges, the Gradex project proposes an automated evaluation system that utilizes Optical Character Recognition (OCR) and Natural Language Processing (NLP) to streamline and standardize answer sheet correction. By replacing the traditional manual evaluation process with AI-driven mechanisms, we can ensure faster, more consistent, and transparent grading.

This system uses Tesseract OCR to extract handwritten or printed text from scanned answer sheets. The extracted text is then preprocessed and evaluated using Ollama language models, which analyze the content against provided answer keys and question papers. This unified evaluation logic ensures fairness and eliminates personal bias.

The project also features a user-friendly interface where teachers can upload answer sheets, organize evaluations by exam type (e.g., CAT 1, Term 1), and access results in real-time. All evaluated data is stored in a structured database and can be exported in CSV format or PDF format for reporting or record-keeping.

1.2 Problem Statement

Despite advancements in teaching technologies, most educational institutions still rely on traditional, manual methods for evaluating exam papers. This manual evaluation process is labor-intensive, time-consuming, and prone to inconsistency due to individual differences among evaluators. Additionally, the lack of standardization in marking schemes can lead to unfair assessments.

To overcome these issues, Gradex introduces an AI-based automated evaluation system. This solution reduces correction time, eliminates evaluator bias, and ensures a consistent and objective grading process. The system also provides quick and accessible results, improving the overall efficiency of academic evaluations.

1.3 Objectives

The objectives of the Gradex system are:

- To automate the evaluation of internal exam papers and reduce the time and manual effort required.
- To ensure consistency in marking by applying a common evaluation standard powered by AI.
- To minimize human errors and emotional bias in the correction process.
- To provide instant results and transparent grading accessible to both teachers and students.
- To categorize and manage results exam-wise (e.g., Term 1, CAT 1, CAT 2) for easier reporting and tracking.

1.4 Scope

The scope of this project is to develop a **fully automated answer sheet evaluation platform** for internal academic assessments. The system performs the following tasks:

- Extracts text from scanned answer sheets using Tesseract OCR.
- Uses Ollama LLMs to compare extracted answers with the answer key.
- Scores answers, calculates total marks, and stores the results in a database.
- Offers an interface for teachers to upload, categorize, and manage answer sheets.
- Provides downloadable CSV reports for easy sharing and record maintenance.
- Can be extended for use in schools, colleges, coaching centers, and competitive exam bodies.

Chapter 2

LITERATURE SURVEY

A literature survey or a literature review in a project report is that section which shows the various analysis and research made in the field of interest and the results already published, taking into account the various parameters and the extent of the project.

2.1 Introduction

In education, grading handwritten exam answers is a time-consuming and labor-intensive task. Traditional grading approaches face challenges, such as subjectivity, inconsistency, and human error, particularly with large volumes of scripts. Automating this process can not only reduce the grading workload but also provide faster and fairer feedback to students.

To address these needs, the proposed system leverages advanced technologies, integrating Optical Character Recognition (OCR) with Large Language Models (LLMs) such as BERT and LLaMA. OCR enables accurate text extraction from handwritten responses, which can then be refined and interpreted by LLMs for effective grading. These models are adept at understanding semantic meaning, ensuring that answers are evaluated based on content rather than exact phrasing. This system also enhances reliability by mitigating OCR errors and improving grading accuracy across varied subjects and response styles.

By combining OCR and LLM-based semantic analysis, this system aims to transform grading into a more efficient, scalable, and objective process. The result is a significant step forward in educational technology, offering a consistent grading solution adaptable to diverse educational environments and large-scale assessments.

2.2 Automated Evaluation of Student Answers using NLP Techniques

Authors: Kumar A., Singh R.

Year: 2020

This research focused on designing a rule-based Natural Language Processing (NLP) system to evaluate descriptive answers submitted by students. The system processes both student responses and model answers through tokenization, stop-word removal, and stemming. It then performs keyword matching and uses cosine similarity to measure textual overlap between the student's answer and the model solution. The evaluation is based on the number of matched keywords and overall similarity. The system provides a lightweight solution for automating descriptive answer scoring in controlled academic environments.

Disadvantages:

- The method lacks semantic understanding and cannot comprehend paraphrased or grammatically restructured answers.
- It performs poorly when students use synonyms or varied sentence structures.
- It requires manual configuration of keywords and threshold similarity values for different questions

2.3 Intelligent Assessment System using Machine Learning for Descriptive Answers

Authors: Sharma N., Patel V.

Year: 2021

This system adopted a data-driven machine learning approach to score descriptive student answers. A training dataset was compiled consisting of various answers labeled with scores assigned by human examiners. Using TF-IDF for feature extraction and classifiers like Support Vector Machines (SVM) and Random Forests, the model was trained to predict scores based on patterns learned from the labeled data.

This approach enabled generalization to new student answers without predefined rules, and allowed faster and automated scoring once trained.

Disadvantages:

- The model's performance was highly dependent on the quality and diversity of the training data.
- It was not effective when applied to questions outside the training distribution or domain.
- The system lacked semantic reasoning and contextual understanding.

2.4 Automatic Grading of Short Answers using BERT

Authors: Zhang Y., Liu H.

Year: 2022

This project introduced a transformer-based model, BERT, to perform automatic grading by capturing the deep contextual meaning of student answers. The model was fine-tuned on a dataset containing question-answer-score triples. BERT's attention mechanism allowed the system to understand the relevance of each word in context, enabling it to evaluate even paraphrased or semantically equivalent answers accurately. This approach showed significant improvement over traditional keyword or rule-based models.

Disadvantages:

- The system was computationally expensive, requiring GPU acceleration for real-time use.
- It required careful fine-tuning for different subjects or types of questions.
- Interpretability was a challenge, as deep learning decisions were opaque to users and teacher

2.5 OCR-Based Automated Examination System

Authors: Das S., Roy A.

Year: 2019

This work focused on using Optical Character Recognition (OCR) to digitize handwritten student answers. The system used Tesseract OCR to convert scanned answer sheets into text. After extraction, the system applied basic keyword-based matching to evaluate content correctness. This project served as an initial step towards paperless evaluations by eliminating manual data entry.

Disadvantages:

- OCR performance degraded significantly with poor handwriting or low image quality.
- Keyword matching lacked depth and failed to consider context or answer coherence.
- The system was not scalable for subjective, multi-paragraph answers.

2.6 Assessment using AI: A Comparative Study of Rule-Based and ML Models

Authors: Thomas J., Mehta R.

Year: 2021

This study compared different AI approaches — rule-based, traditional machine learning, and deep learning models — for automated evaluation. The authors developed small-scale prototypes for each category and analyzed them for accuracy, training time, scalability, and generalization. This comparative framework helped highlight the strengths and limitations of each technique, offering guidance for hybrid model development.

Disadvantages:

- No single model performed well across all question types.
- Deep models required significant resources and domain-specific tuning.
- Rule-based systems needed continuous manual maintenance and updates.

Chapter 3

System Analysis

3.1 Existing System

3.1.1 Traditional Manual Grading and Its Limitations

Manual evaluation of descriptive answer sheets remains the dominant approach in most educational institutions. In this process, teachers individually review each student's handwritten responses, compare them with expected answers, and assign marks based on their interpretation and experience. While this method allows for human judgment and flexibility, it suffers from several critical limitations. Firstly, it is extremely **time-consuming**, especially when dealing with large volumes of answer sheets, such as during term-end or competitive examinations. Teachers often spend long hours reading through pages of handwritten content, leading to fatigue and slower result processing.

Another key drawback is the **subjectivity and human bias** involved in manual grading. Different teachers may interpret and score the same answer differently due to personal preferences, expectations, or unconscious bias. This lack of standardization results in inconsistent and unfair scoring. Additionally, **errors and inconsistencies** frequently occur, particularly under tight deadlines. Mistakes in mark calculation, misinterpretation of answers, and overlooked content are common when teachers are under pressure or when answer sheets are not clearly written. These challenges have highlighted the urgent need for a more objective, scalable, and accurate solution for evaluating descriptive answers.

3.1.2 Automated Grading for Objective-Type Questions

Automated grading systems have already seen success in handling objective-type assessments like multiple-choice questions, true/false, and fill-in-the-blank formats. E-learning platforms such as Google Forms, Moodle, and Edmodo employ rule-based grading mechanisms, which rely on exact answer matching. These systems offer speed and accuracy in assessing large datasets with zero ambiguity in answers. However, their limitations become apparent when used to evaluate descriptive or subjective answers, as rule-based systems struggle to interpret synonyms, paraphrased responses, or sentence-level meaning. They are unable to provide meaningful feedback or assess the creativity and reasoning skills exhibited in longer answers.

3.1.3.OCR-Based Answer Extraction

To digitize and process handwritten answer sheets, Optical Character Recognition (OCR) is an essential technology. Several tools and APIs like Tesseract OCR, Google Vision API, and OpenAI's OCR models have been applied to extract text from scanned images or photographs of answer sheets. While these tools work well for printed text, handwritten content remains a major challenge due to varied writing styles, poor scan quality, image noise, and inconsistencies in character formation. These limitations often result in partially or incorrectly extracted text, which negatively affects the downstream evaluation process. The current study overcomes this by employing Ollama OCR, a deep learning-based handwriting recognition model that improves accuracy significantly. It leverages neural networks trained on diverse handwriting samples, enabling better generalization across different student answer scripts.

3.1.4.NLP and AI in Descriptive Answer Evaluation

Recent advancements in Natural Language Processing (NLP) have opened the door to AI-powered evaluation systems that go beyond keyword matching and instead focus on understanding the semantic meaning of text. By applying word embeddings like Word2Vec, GloVe, and FastText, along with more advanced transformer-based models such as BERT, T5, LLaMA, and GPT, these systems can analyze context and evaluate the logical structure and relevance of student answers. Unlike traditional rule-based systems, NLP models are capable of context-aware assessment, recognizing correct answers written in different ways using synonyms or paraphrased structures.

One of the major strengths of AI-driven evaluation is semantic similarity matching, where student responses are compared with model answers not through exact word matching but through contextual alignment. These models also handle grammatical variations and sentence restructuring, thus offering a more flexible and accurate evaluation method. Studies show that transformer-based models like BERT and GPT consistently outperform traditional techniques in grading accuracy and interpretation of complex answers, making them suitable for deployment in educational settings.

3.2 Disadvantages of Existing Systems

- Time-consuming process for teachers, especially during large-scale examinations.
- Lack of scalability with increasing student population.
- High risk of human bias and subjectivity in grading.
- Inconsistencies and errors due to fatigue or oversight.
- Minimal or no feedback provided to students for improvement.
- Poor performance of automated systems on descriptive answers.
- Dependence on exact keyword matching in automated grading systems.
- OCR tools struggle with handwritten text due to variations in writing style and image quality.
- No digital storage or centralized access to student performance data.
- Limited support for semantic understanding and context-aware evaluation.
- Absence of data-driven analytics and feedback loops for institutional improvement.

3.3 Proposed System

The proposed system, Gradex, is an AI-powered answer sheet evaluation platform designed to overcome the limitations of traditional manual grading and rule-based automated systems. It combines advanced OCR, NLP, and semantic analysis techniques to accurately evaluate descriptive student answers. By digitizing handwritten responses, extracting meaningful content, and applying intelligent scoring mechanisms, Gradex aims to provide faster, more consistent, and unbiased evaluation of academic answer sheets.

3.3.1 Handwritten Text Extraction Using Ollama OCR

At the core of the system is the use of Ollama OCR, a deep learning-based optical character recognition engine designed to accurately read and extract text from handwritten documents. Unlike conventional OCR tools like Tesseract or Google Vision API, which struggle with varied handwriting styles, Ollama OCR leverages advanced neural networks to process noisy, distorted, or complex handwriting. This enables the system to accurately digitize handwritten responses from scanned answer sheets, forming a reliable input for further analysis.

3.3.2 Text Preprocessing and Cleaning

Once the text is extracted, it undergoes a robust text preprocessing pipeline to clean and normalize the content. This step includes removing unwanted characters, correcting common OCR errors, and structuring the response into coherent sentences. Preprocessing also involves tokenization and part-of-speech tagging, which prepares the text for semantic analysis by ensuring clarity and structure in the response data.

3.3.3 Semantic Evaluation Using NLP Models

The core evaluation mechanism of Gradex uses state-of-the-art NLP models to assess the content of student answers. These models—such as BERT, GPT, and LLaMA—are capable of understanding semantic meaning rather than relying on keyword matches. The system compares each student answer with the corresponding model answer (key) using semantic similarity scoring, which considers synonyms, paraphrasing, and logical structure.

This allows the system to grade answers that are correct but differently worded, a capability absent in traditional systems. The models also account for grammatical variations and logical coherence, making the grading process more holistic and context-aware.

3.3.4 Score Calculation and Feedback Generation

Based on the semantic similarity scores and logical structure of the response, Gradex assigns marks to each question. It uses a configurable rubric system to ensure flexibility across different subjects and question types. In addition to scores, the system generates automated feedback highlighting the strengths and weaknesses of the student's response. This feedback helps students understand their mistakes and improve in future assessments

3.3.5 Teacher Dashboard, Student Information, and Result Accessibility

Gradex provides a comprehensive teacher dashboard that enables educators to efficiently manage the entire evaluation workflow. Teachers can upload scanned answer sheets, create and manage various exams such as CAT 1, CAT 2, and Term 1, and monitor detailed student performance analytics. The dashboard offers valuable insights into individual and class-wide progress, including average scores, frequently missed concepts, and improvement trends. Additionally, the system organizes student results under exam-wise categories, making it easy to retrieve and review historical performance. When a teacher accesses the "Student Info" section, they can view all attempted exams by a student along with the respective scores and AI-generated feedback. Teachers can also review and, if necessary, override AI-evaluated answers, ensuring a transparent and flexible grading process that supports both automation and human oversight.

3.4 System Architecture

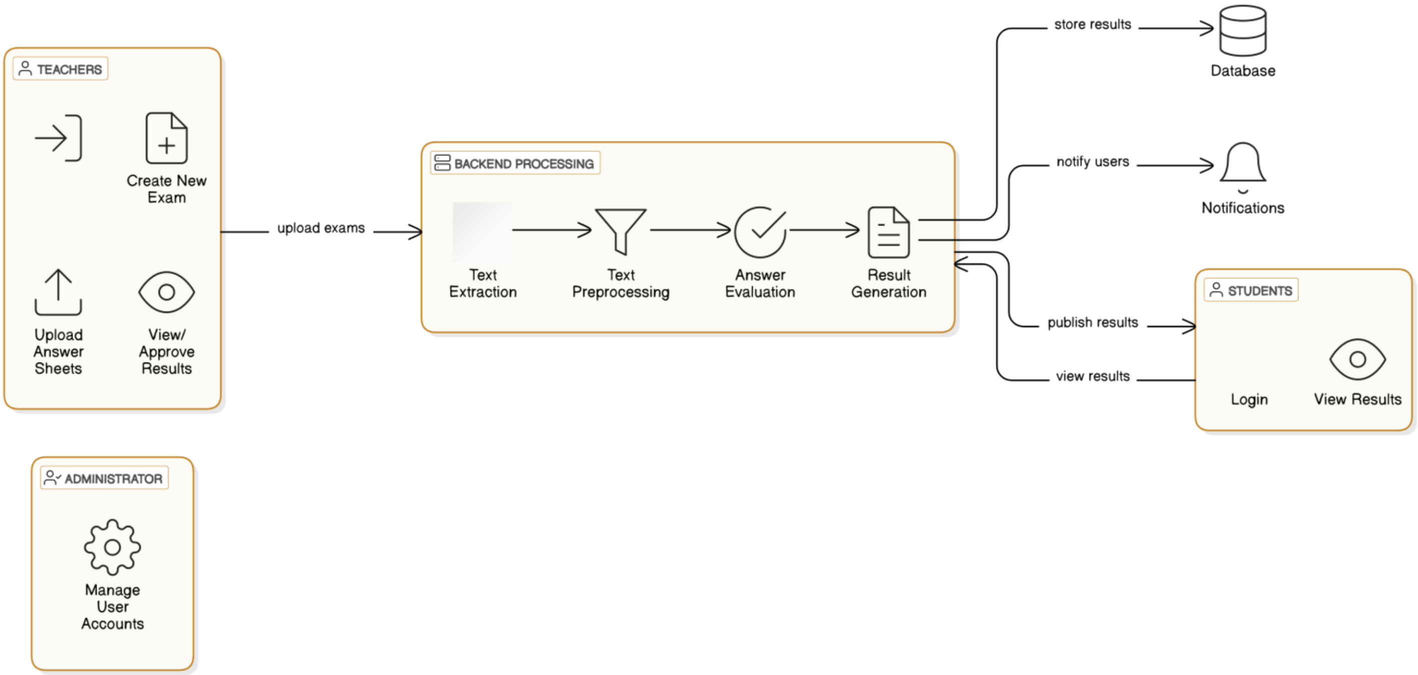

Fig: 3.1 GradeX Architecture

Chapter 4

System Requirements

4.1 Overall Description

This section describes the **system requirements** for the Gradex project, focusing on both **hardware** and **software** components needed to run the system effectively. These requirements are essential for ensuring that the software functions efficiently and reliably within the specified environment.

4.2 Specific Requirements

4.2.1 Hardware Requirements

The hardware requirements for the Gradex system outline the necessary physical components to run the software smoothly. The system should be capable of processing large batches of scanned answer sheets and executing complex AI models. The recommended hardware specifications are:

- **Processor:** A **2.0 GHz** or higher multi-core processor (such as Intel i5 or equivalent). This is required to handle the processing tasks of OCR and NLP models efficiently.
- **RAM:** A minimum of **4 GB of RAM** is recommended to allow smooth operation, especially when handling large batches of answer sheets. More memory may be required for large-scale deployments.
- **Disk Space:** The system will need **1 GB** or more of free hard disk space to store the software, processing data, and results. This space is also used for caching images and saving evaluation results.
- **Scanner:** A high-resolution scanner (at least **300 DPI**) is required to ensure that the scanned images of the answer sheets are clear and suitable for OCR processing.

- **Network Connection:** A stable internet connection is required for cloud-based functionalities, such as accessing AI models, uploading files, and synchronizing data with the server.

4.2.2 Software Requirements

The **software** requirements for Gradex are the essential tools and technologies needed to build, run, and maintain the system. These requirements include the operating system, programming languages, frameworks, and development tools.

- Operating System: Windows 10, Linux (Ubuntu), or macOS
- Programming Language: Python 3.9 or above
- Web Framework: Django
- Database: PostgreSQL (recommended) or SQLite (for testing)
- OCR Tool: Tesseract OCR model
- Evaluation Models: Ollama LLMs for semantic answer evaluation
- Development Tools: VS Code or PyCharm
- Model Testing: Google Colab Notebook
- Deployment: Pythonanywhere.com

4.3 Functional Requirements

- The system must allow teachers to log in securely and manage exams (e.g., CAT 1, CAT 2, Term 1).
- Teachers should be able to upload scanned student answer sheets in image format.
- The system must extract text from uploaded images using Ollama OCR.
- Extracted text should be evaluated by comparing it with key answers using Ollama LLMs.
- The system must automatically assign marks based on semantic similarity.
- Teachers should be able to review and override AI-evaluated scores.

- The system must store and categorize student results based on exam type.
- The system should provide detailed performance reports for each student and exam.
- Teachers should be able to search and view individual student performance history.

4.4 Non-Functional Requirements

- **Reliability**: The system should perform evaluations accurately and consistently under normal load.
- **Scalability**: It should support increasing numbers of users and answer sheets without performance degradation.
- **Security**: Only authenticated teachers can access and modify student data. Sensitive information must be protected.
- **Performance**: OCR and evaluation processes should complete within an acceptable response time.
- **Maintainability**: The codebase should be modular and well-documented to ease future updates and debugging.
- **Usability**: The interface should be intuitive and accessible for teachers with basic technical skills.
- **Availability**: The system should be available online with minimal downtime for maintenance or updates.

Chapter 5

System Design

5.1 Architectural Diagram

The architectural design gives the description about the overall system design. It is specified by identifying the components defining the control and data flow between them. The arrow indicates the connection and rectangular box represents the functional units. The Fig. 5.1 shows the architectural diagram of this project which shows the overall operation from uploading question papers to the evaluation of the student answer scripts.

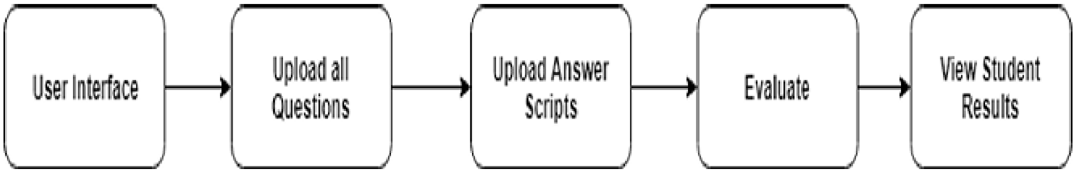

Fig:5.1 GradeX Flow Diagram

5.2 UML Diagram

5.2.1 Use Case Diagram

A Use Case Diagram is a behavioral diagram in UML (Unified Modeling Language) that visually represents the interactions between users (actors) and the system, showing the system's functional requirements.

In the context of the Gradex project, the use case diagram identifies the core functionalities of the AI-based evaluation system and maps out the roles of its primary users—teachers and students.

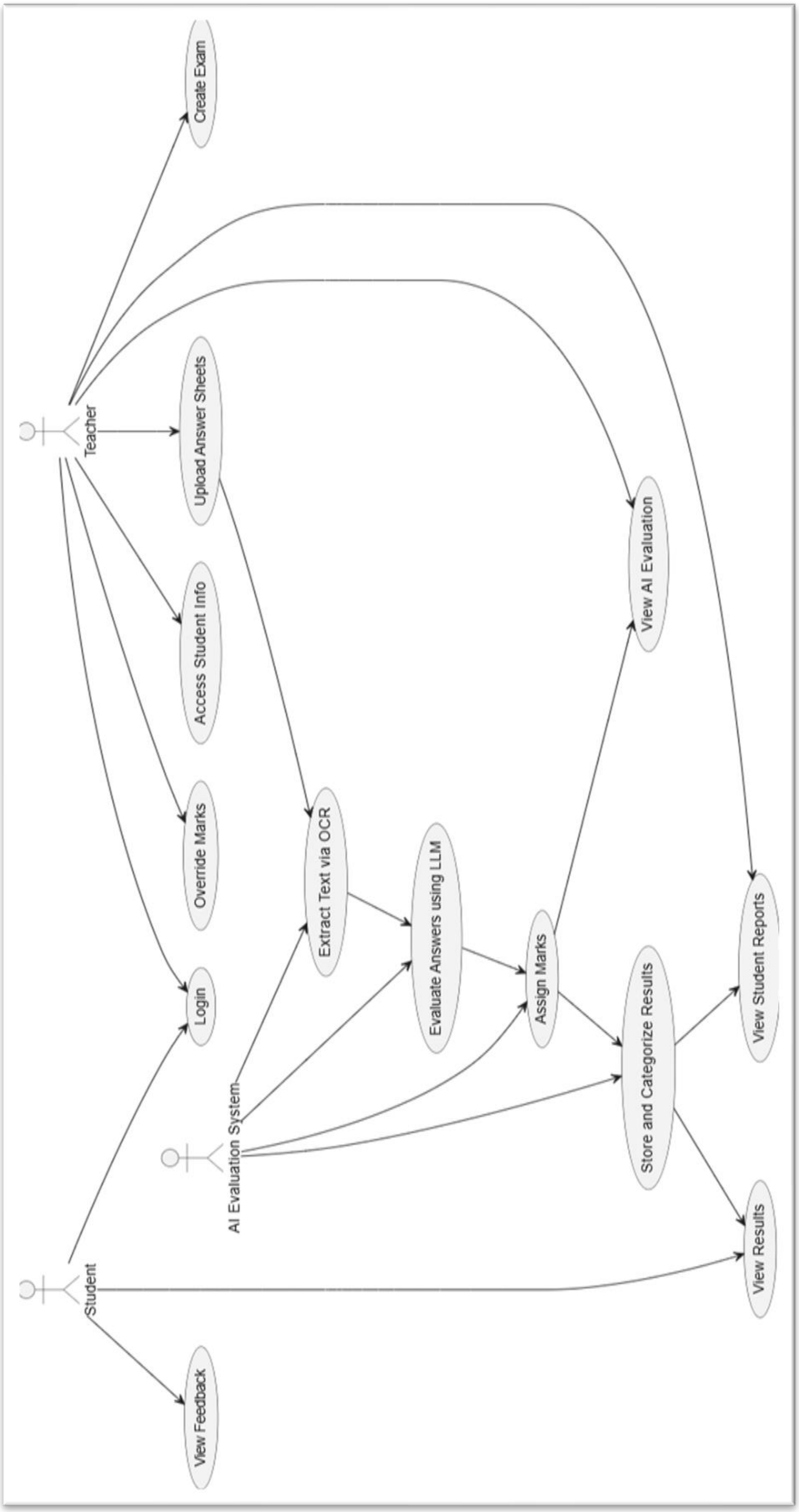

Fig 5.2 Use Case Diagram

5.2.2 Class Diagram

A Class Diagram is a type of static structure diagram in the Unified Modeling Language (UML) that describes the structure of a system by showing its classes, attributes, methods, and the relationships among the classes. It serves as a blueprint of the system's object-oriented design and forms the foundation for coding and system implementation.

In the context of the Gradex project, the class diagram models the key components of the AI-based answer sheet evaluation system. It includes entities such as Student, Teacher, Exam, AnswerSheet, and EvaluationResult, each represented as a class with defined attributes (e.g., student ID, exam name) and methods (e.g., uploadSheet(), evaluateAnswer()). The relationships between these classes—such as association, aggregation, or inheritance—are also illustrated to show how different components of the system interact.

Class diagrams help developers and stakeholders understand the internal structure and logic of the application. They are essential during the design phase for identifying objects, organizing code, and ensuring the system supports scalability, maintainability, and reusability.

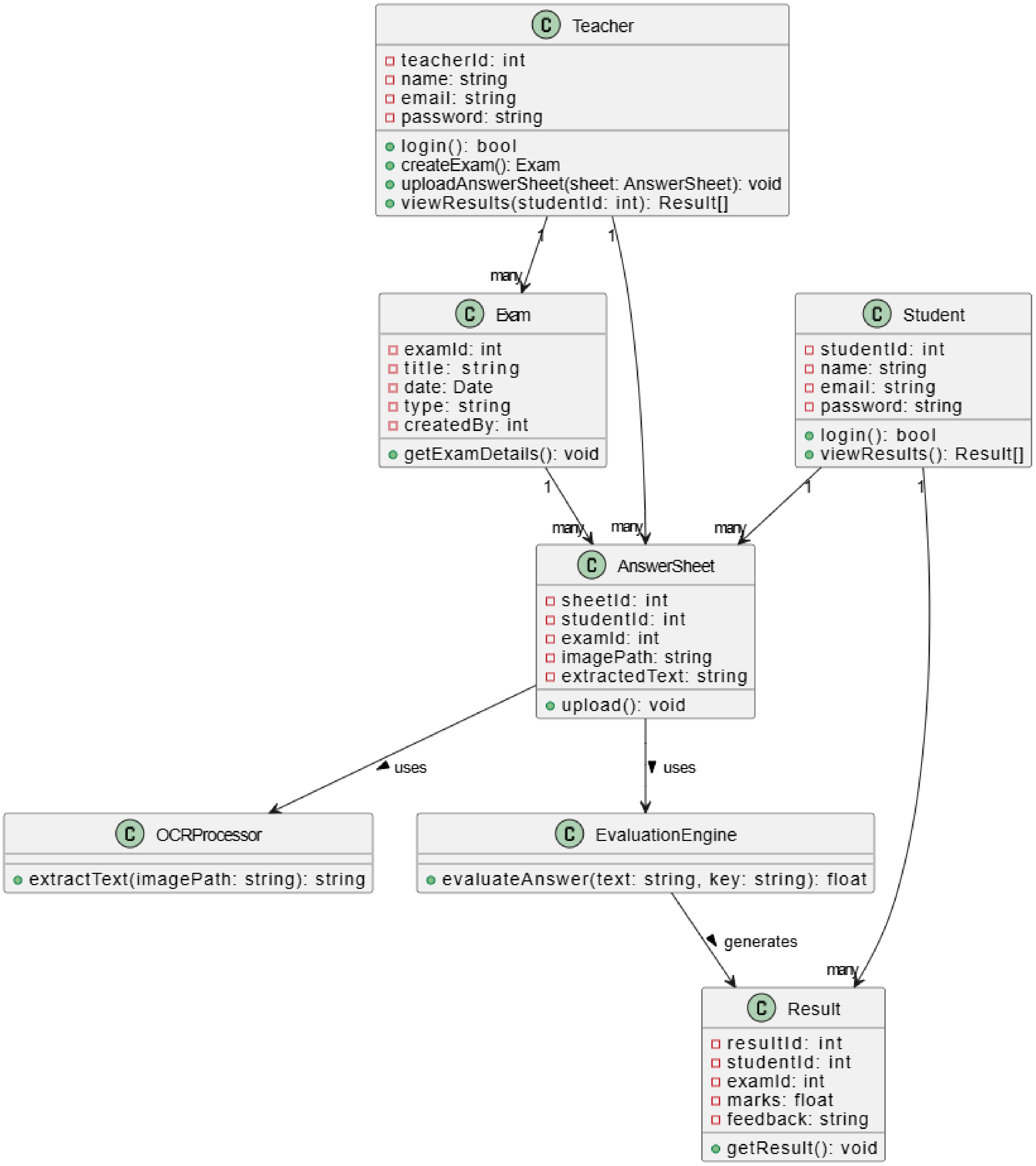

Fig 5.3 Class Diagram

5.2.3 Sequence Diagram

A Sequence Diagram is a type of interaction diagram in the Unified Modeling Language (UML) that illustrates how objects in a system interact with one another over time. It focuses on the order of messages exchanged between various system components (objects or actors) to accomplish a specific task or use case.

In the context of the Gradex project, the sequence diagram models the flow of interactions between the key actors—Teacher, Student, and the Gradex system components such as the OCR engine, evaluation module, and database. It visually represents how, for example, a teacher uploads an answer sheet, the system extracts text using OCR, evaluates the content using AI models, and stores the results. Similarly, it shows how a student can later access the evaluated results.

The sequence diagram emphasizes temporal ordering, showing which interactions happen first and how control flows among the system parts. It is crucial for understanding runtime behavior, debugging logic flow, and aligning system operations with user expectations.

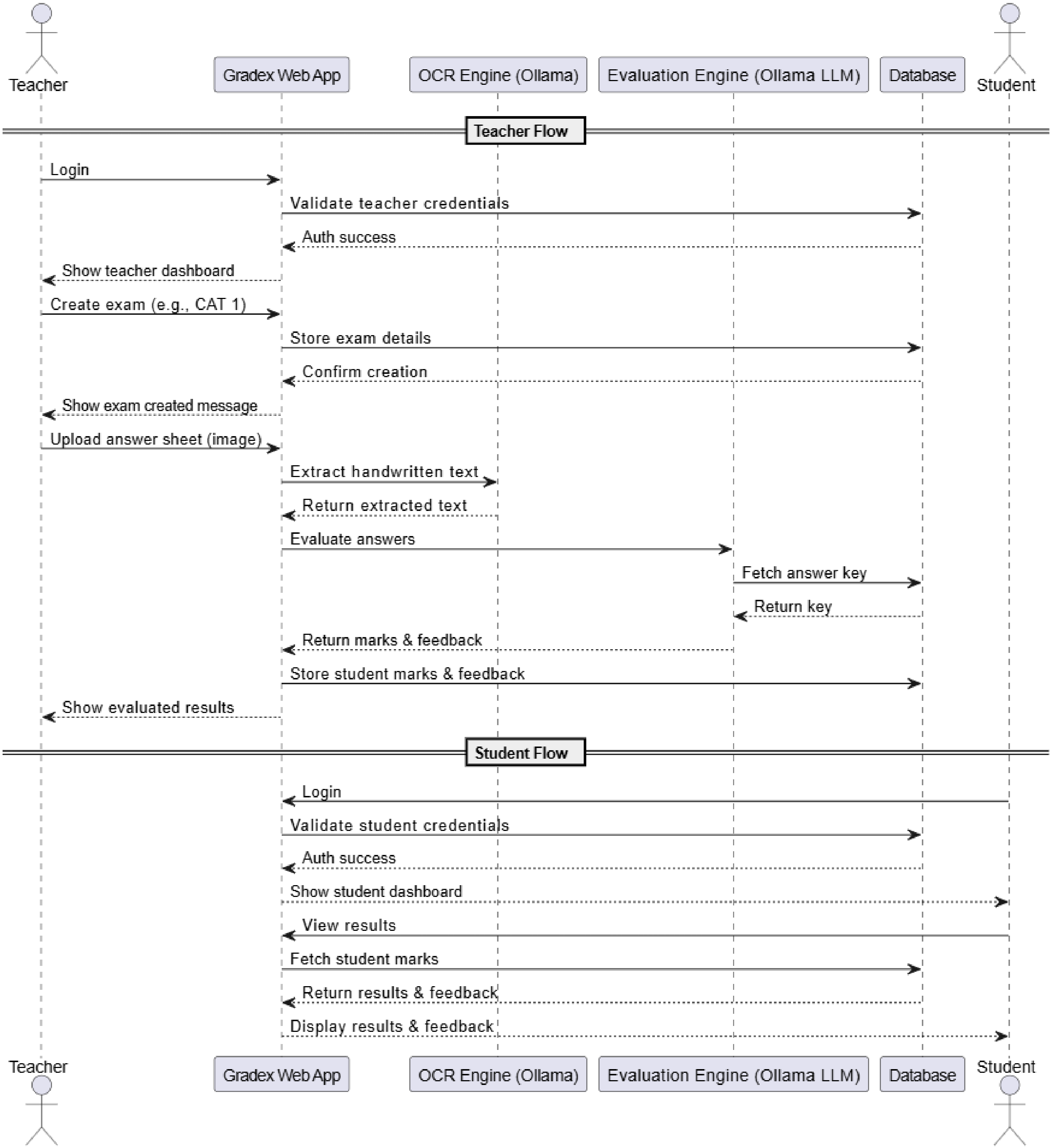

Fig 5.4 Sequence Diagram

5.2.4 Activity Diagram

An Activity Diagram is a type of behavioral diagram in the Unified Modeling Language (UML) that represents the workflow or sequence of activities in a system. It focuses on dynamic aspects of the system by modeling the flow of control from one activity to another, including decisions, parallel processes, and loops.

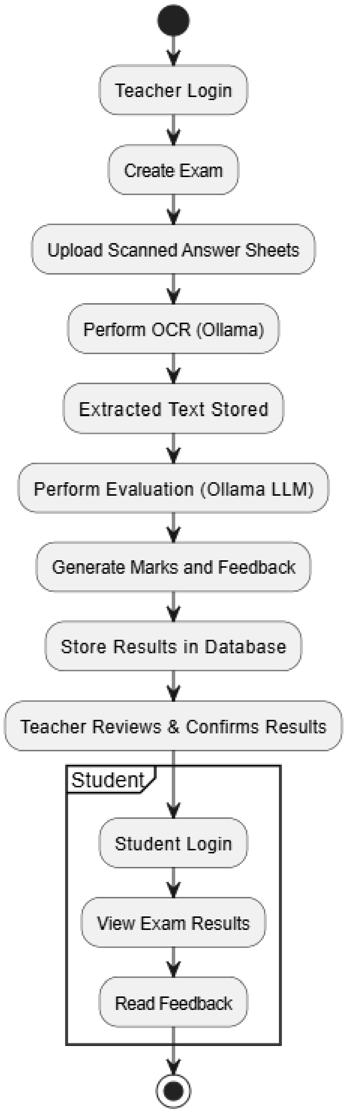
Teacher Login
Create Exam
Upload Scanned Answer Sheets
Perform OCR (Ollama)
Extracted Text Stored
Perform Evaluation (Ollama LLM)
Generate Marks and Feedback
Store Results in Database
Teacher Reviews & Confirms Results
Student
Student Login
View Exam Results
Read Feedback

Fig 5.5 Activity Diagram

5.2.5 DFD Diagram

A **Data Flow Diagram (DFD)** is a graphical representation used to visualize how **data moves through a system**. It shows the flow of information between various **processes**, **data stores**, **external entities**, and the system itself. DFDs are widely used during the system analysis phase to understand the functional aspects of a system and how it handles data.

In the context of the Gradex project, a DFD illustrates how data such as scanned answer sheets, extracted text, evaluation results, and student records flow between key components. These include external entities like Teachers and Students, internal processes like OCR extraction and AI-based evaluation, and data stores such as exam records and student databases.

The DFD helps in identifying the inputs, outputs, and transformations of data within the system. It ensures that all data interactions are accounted for, helping developers and analysts spot bottlenecks, redundancies, or missing components in the system design.

Gradex's DFD provides a logical view of the system without focusing on implementation details, making it easier to communicate system functionality with both technical and non-technical stakeholders.

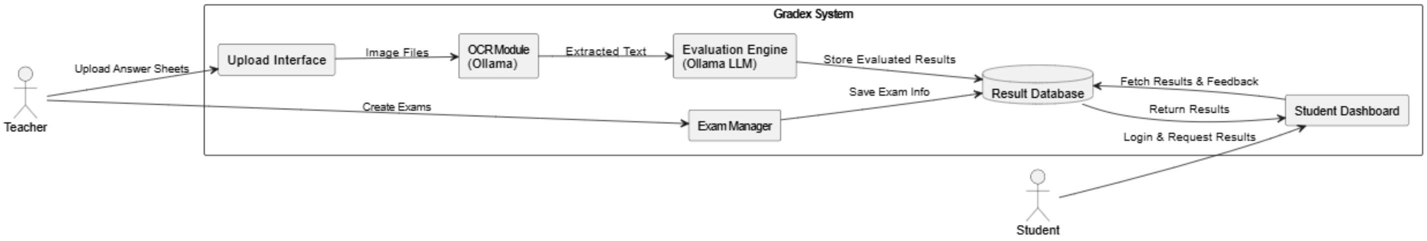

Fig 5.6 DFD Diagram

5.2.6 ER Diagram

An **Entity-Relationship (ER) Diagram** is a data modeling technique used to visually represent the **entities** within a system and the **relationships** between them. It provides a high-level conceptual view of how data is structured and connected in a database, forming the foundation for designing relational databases.

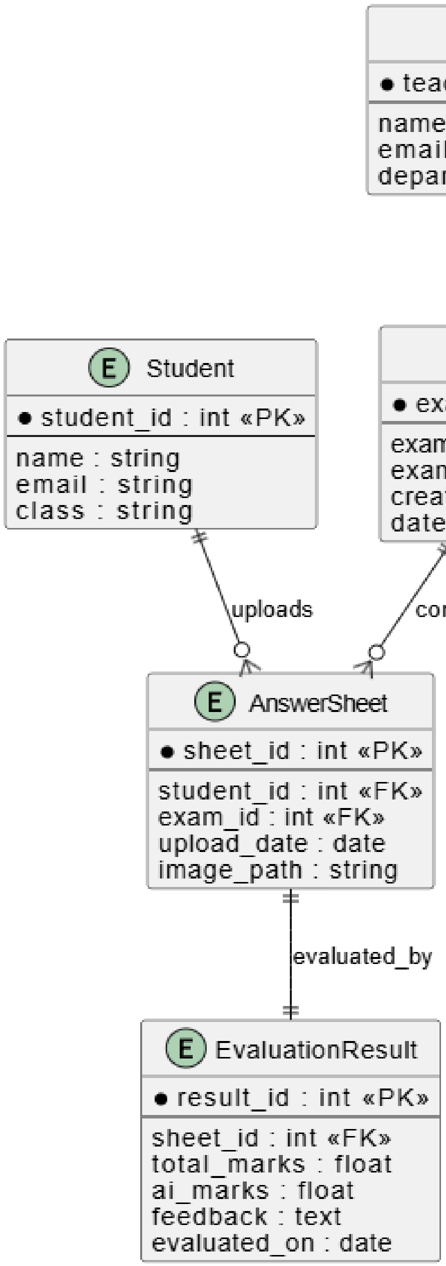

Fig 5.7 ER Diagram

Chapter 6

Modules Implementation

6.1 Modules List

- Tesseract OCR Module
- Text Preprocessing Module
- Question Answer Mapping Module
- Ollama 3.1 Text Refinement Module
- BERT Semantic Matching Module
- Grading and Feedback Module
- Exam Module
- ExamEvaluation Module
- Teacher Dashboard Module
- Student Dashboard Module
- Authentication Module

6.2 Tesseract OCR Module

The Tesseract OCR Module is the entry point of the Gradex evaluation system. Its main responsibility is to extract raw text from scanned student answer sheets. This is essential for automating the evaluation process, as all further processing depends on accurate text extraction.

6.2.1 Image Input Handling

This module accepts various formats like JPG, PNG, and PDF. These files typically contain handwritten or printed answers submitted by students. High-resolution images yield better OCR accuracy.

6.2.2 Preprocessing Techniques

Before passing the image to the Tesseract engine, the system applies several preprocessing steps to enhance clarity and readability:

- Grayscale conversion to simplify color data.
- Thresholding to convert the image to binary (black and white).
- Noise reduction using blurring or morphological operations.

These steps help reduce OCR errors caused by shadows, smudges, or low-quality scans.

6.2.3 Tesseract Integration

The core of this module is powered by the Tesseract OCR engine. It reads the preprocessed images and converts them into machine-readable text. Tesseract supports multilingual recognition and layout analysis, making it suitable for a variety of answer sheet formats.

6.2.4 Text Output

The final output of this module is plain, unstructured text that represents the student's written responses. This output is passed directly to the Text Preprocessing Module for further cleaning and structuring before evaluation

6.3 Text Preprocessing Module

The Text Preprocessing Module is responsible for cleaning and structuring the raw text extracted by the Tesseract OCR Module. Since OCR output can be noisy, inconsistent, or contain formatting issues, this module ensures the text is usable for accurate question-answer mapping and evaluation.

6.3.1 Purpose

The main goal is to transform unstructured, messy OCR text into clean, well-organized content that resembles how a human would write or interpret it. This includes removing unwanted characters, fixing sentence structures, and identifying sections relevant to each question.

6.3.2 Cleaning and Normalization

This step involves a series of text-cleaning techniques, such as:

- Removing special symbols, extra whitespace, or page numbers.
- Converting all characters to a consistent case (e.g., lowercase).
- Fixing common OCR errors like misrecognized characters (e.g., ‘0’ instead of ‘O’, ‘1’ instead of ‘I’).
- Standardizing punctuation.

6.3.3 Segmentation and Structuring

The module splits the cleaned text into logical segments like:

- Individual answers or paragraphs.
- Question number identification (e.g., “Q1”, “1.”, etc.).
  This helps in aligning each answer to the corresponding question during the mapping phase.

6.3.4 Spell Correction and Grammar Fixing

Lightweight spelling correction and grammar enhancement may also be applied using NLP techniques or LLMs (like Ollama) to improve the language quality before semantic comparison.

6.3.5 Output

The output is a cleaned, structured text format—typically in the form of a list or dictionary where each entry maps to a specific question. This structured data is then forwarded to the **Question-Answer Mapping Module** for further alignment and evaluation

6.4 Question-Answer Mapping Module

The Question-Answer Mapping Module is essential for organizing and linking the student's responses to their corresponding exam questions. Once the text has been cleaned and preprocessed, the system faces the challenge of accurately identifying which portion of the text answers which question. This can be difficult because students may not always number or format their answers clearly, and OCR errors can sometimes cause further confusion. To address this, the module employs a combination of pattern recognition techniques and semantic analysis. Pattern recognition detects common markers like "Q1," "1.", or other numbering styles that indicate the start of an answer. Semantic analysis uses language models and contextual understanding to interpret the content and determine the most probable question a particular answer relates to, even if explicit numbering is missing or ambiguous. The module also references the official question paper or answer key structure as a guide to improve accuracy in alignment. By creating this precise mapping, the module ensures that each student's response is correctly paired with the intended question, which is critical for fair and accurate automated grading. The result is a structured dataset that pairs question identifiers with the corresponding student answers, ready for the subsequent evaluation and grading processes.

6.5 Ollama 3.1 Text Refinement Module

Once raw text is extracted using the OCR module, it is passed to the Ollama 3.1 Text Refinement module for advanced preprocessing. This deep learning-based module is designed to clean and normalize the extracted content. It corrects OCR mistakes such as misrecognized characters, missing punctuation, and misplaced words. Furthermore, it restructures the sentences for coherence and improves readability, ensuring that the text closely resembles a logically written response. This refined output is crucial for the success of downstream NLP tasks, especially semantic evaluation. The module uses 28 language understanding techniques to preserve the original meaning while improving the quality and format of the text.

6.5.1 ALGORITHM STEPS

Step 1: Get Raw OCR Output

- After the answer sheet is scanned and processed using **Tesseract OCR**, the output is often messy:
    - Spelling errors
    - Broken grammar
    - Wrong punctuation
    - Incomplete sentence structure

Step 2: Clean the Basic Noise

- Remove unwanted symbols, extra spaces, page numbers, or unreadable characters.
- Convert everything to lowercase (optional).
- This helps the LLaMA model focus only on the actual content.

Step 3: Prepare the Instruction

- You give LLaMA 3.1 a **clear instruction** like:

"Correct the spelling and grammar of this answer while keeping the original meaning."

- Follow it with the raw student answer from OCR.

Step 4: Run through LLaMA 3.1 (Ollama)

- The instruction and OCR text are sent to **LLaMA 3.1**, which is a powerful large language model.
- It uses:
  - Deep transformer layers
  - Attention mechanisms
  - Language prediction capabilities
- It understands what the student was trying to say and rewrites it in correct English.

Step 5: Receive Refined Answer

- LLaMA returns the corrected version:
  - Sentences are now clear, grammatically correct, and properly structured.
  - It looks like a well-written human answer.

Step 6: Structure the Answers

- Identify which part belongs to which question (Q1, Q2, etc.).
- This is done using clues like "1.", "Q1", or by checking logical breaks.
- Store each cleaned answer separately and neatly.

Step 7: Send to BERT for Grading

- Now that the student answers are cleaned and readable:
  - Pass them to **BERT**

    - BERT will now compare them with the model answer using semantic meaning (not exact words)

Step 8 (Optional): Use LLaMA for Feedback

- You can also use LLaMA 3.1 to generate feedback like:
    - "Well explained."
    - "You missed the definition part."
    - "Use technical terms for better clarity."

6.6 BERT Semantic Matching Module

The BERT Semantic Matching Module is responsible for understanding and comparing the meaning of the student's refined answers with the model or ideal answers provided in the answer key. Using BERT, a powerful transformer-based language model, this module goes beyond simple keyword matching by capturing the context, semantics, and nuances of the language used. It converts both the student's response and the model answer into high-dimensional embeddings that represent their meaning, then calculates similarity scores between these embeddings. This allows the system to evaluate whether the student's answer conveys the correct concepts, even if phrasing or word choices differ significantly from the model answer. By leveraging BERT's deep language understanding capabilities, this module provides a more accurate and fair semantic comparison, which is crucial for assigning meaningful grades and feedback based on content relevance rather than just exact text matches.

6.6.1 ALGORITHM STEPS

STEP 1: Input Preparation

- **Student Answer** and **Answer Key** are taken as input.
- Both are converted into a format BERT can process:
    - [CLS] student_answer [SEP] answer_key [SEP]

- Tokenization is done using **WordPiece Tokenizer**, which breaks text into subword units.

STEP 2: Token Embedding

- Each token is converted into three types of embeddings:
  1. **Token Embedding** – actual word pieces
  2. **Segment Embedding** – distinguishes student answer (Segment A) and answer key (Segment B)
  3. **Position Embedding** – encodes position of each token in the sentence

STEP 3: Feeding into BERT Model

- These embeddings are fed into BERT's **Transformer architecture**:
  - Multiple layers of **self-attention** and **feed-forward networks**
  - BERT learns the contextual relationship between words in both student answer and answer key

STEP 4: Extracting the Output Embedding

- The output vector for the [CLS] token (first token) is used as the **summary representation** of the semantic similarity between the two texts.

STEP 5: Similarity Scoring

- The [CLS] vector is passed through a **fully connected dense layer** and a **sigmoid or softmax** activation function.
- Output: a **semantic similarity score** (0 to 1 or 0 to 100).

STEP 6: Grading Logic

- Based on the similarity score:
  - If score ≥ threshold (e.g., 0.85): **Full marks**
  - Score between thresholds: **Partial marks**
  - Score < lower threshold: **Low or zero marks**

STEP 7: Feedback Generation

- Along with the score, your system (GradeX) uses the difference in embeddings to generate:
    - Feedback on missing content
    - Suggestions for improvement

6.7 Grading and Feedback Module

The Grading and Feedback Module is the core component that assigns scores to student answers based on the semantic similarity results and predefined grading criteria. After the BERT Semantic Matching Module evaluates how closely the student's answers align with the ideal responses, this module applies specific rubrics and thresholds to determine the appropriate grade or marks for each answer. It considers factors such as completeness, relevance, and correctness derived from the semantic similarity scores. In addition to numeric grading, the module generates detailed feedback to help students understand their mistakes, strengths, and areas for improvement. This feedback may include explanations, suggestions, or references to related study materials. By automating the grading and feedback process, this module not only saves teachers significant time but also provides students with consistent and objective evaluations, enhancing the learning experience.

6.8 Teacher Dashboard Module

The Teacher Dashboard Module serves as a centralized control panel designed specifically for educators to streamline and manage all aspects of the examination and evaluation process. Through this module, teachers can upload exam materials such as question papers and scanned student answer sheets, initiating the automated grading workflow seamlessly. It offers robust functionalities for organizing exams, allowing teachers to create new exams, define exam types (like term exams or periodic assessments), and set grading criteria or rubrics tailored to each exam's requirements.

Once the grading process is complete, the dashboard presents comprehensive insights into student performance. Teachers can view individual student profiles, detailed answer evaluations, scores, and generated feedback, helping them to quickly identify strengths and weaknesses at both the student and class levels. The module also supports exam-wise result categorization and trend analysis over time, enabling teachers to monitor academic progress and intervene where necessary.

Additionally, the Teacher Dashboard facilitates communication by allowing teachers to annotate reports or add personalized comments before sharing results with students. It includes options to export reports in various formats for official record-keeping or further review. The interface is designed to be intuitive and accessible, reducing administrative workload and improving accuracy, thereby empowering educators to focus more on teaching and student development.

6.9 Student Dashboard Module

The Student Dashboard Module is the interface designed for students to access their exam results, detailed feedback, and progress reports in a clear and user-friendly manner. Once teachers upload and grade answer sheets, the processed results are made available through this module, allowing students to review their performance on various exams categorized by exam types such as term exams, class assessments, or tests.

This module provides a personalized view where students can see their scores alongside detailed comments and suggestions generated by the grading system. These insights help students understand their strengths, pinpoint specific mistakes, and receive recommendations for improvement. The dashboard may also include visual elements like charts and graphs to track performance trends over time, fostering self-awareness and motivation.

Additionally, the Student Dashboard may offer features for students to download their reports or request clarifications from teachers, creating a two-way communication channel. By giving students transparent and timely access to their academic data, this module supports a more engaged and informed learning experience, encouraging continuous improvement and active participation in their education journey.

6.10 Exam Module

The Exam Module is responsible for storing and managing all exam-related data within the system. It maintains comprehensive records of exam details, including the exam name, type (such as term exams, CATs, or quizzes), dates, duration, and associated question papers. This centralized repository ensures that all exam metadata is well-organized and easily accessible for processing and reference throughout the grading workflow.

The module supports the creation, updating, and deletion of exam records, allowing administrators or teachers to manage multiple exam sessions efficiently. It also links each exam to its corresponding question paper and answer key, which are critical for accurate evaluation. By maintaining exam data systematically, this module enables smooth integration with other system components, such as the Question-Answer Mapping and Exam Evaluation Modules, ensuring consistent and reliable assessment processes.

6.11 Exam Evaluation Module:

The Exam Evaluation Module is responsible for storing and managing the corrected answer data and evaluation results for each exam. After the grading process, this module captures detailed records of each student’s scores, answer-level feedback, and overall performance metrics. It maintains a structured database that links corrected answers to the corresponding exam, student, and question, enabling easy retrieval and analysis.

This module also supports tracking evaluation history, allowing teachers and administrators to review past results, monitor academic progress, and generate reports over different assessment periods. By organizing and preserving evaluation data systematically, the Exam Evaluation Module ensures transparency, accuracy, and accountability in the grading process, while providing a solid foundation for generating insights and enhancing educational outcomes.

6.12 Authentication Module

The Authentication Module manages user access and security within the system. It handles the registration, login, and role-based authorization processes for different users such as teachers, students, and administrators. By implementing secure authentication protocols, this module ensures that only authorized individuals can access sensitive exam data, upload answer sheets, or view evaluation results.

It supports features like password encryption, session management, and possibly multi-factor authentication to protect user accounts and maintain data privacy. Role-based access controls restrict functionalities according to user types, for example, allowing teachers to manage exams and grade papers while students can only view their own results. The Authentication Module is fundamental to maintaining the integrity and confidentiality of the system, safeguarding both academic data and personal information.

Chapter 7

Experimental Results

The experimental results in the **Gradex System** focus on evaluating the performance and accuracy of each of the core modules: the **Tesseract OCR Module**, **Ollama 3.1 Text Refinement Module**, **BERT Semantic Matching Module**, and the overall grading system. The primary goals of these experiments are to measure the effectiveness of text extraction, refinement, semantic matching, and the accuracy of the final grading and feedback process.

7.1 Tesseract OCR Module Performance

The Tesseract OCR module is tested on multiple sets of handwritten and printed answer sheets to measure its accuracy in converting scanned images to machine-readable text. The key performance metrics for this module are:

- **Accuracy of Text Extraction**: The percentage of correctly recognized characters, words, and sentences.
- **Error Rate**: The rate at which Tesseract misidentifies characters or fails to recognize certain parts of the image.
- **Processing Time**: The time taken for Tesseract to process an image and generate text.

Results:

- **Handwritten Text**: When tested on handwritten answer sheets, Tesseract achieved an accuracy rate of approximately 85%, with errors primarily arising from illegible handwriting and distorted images. The error rate was higher when the handwriting quality was inconsistent, with misrecognized characters being the most frequent error.

- **Printed Text**: The OCR performance was significantly better on printed text, with an accuracy rate of about 95%. Errors were minimal and typically related to image quality issues, such as blurriness or low resolution.
- **Processing Time**: The average processing time per page of text (roughly 1-2 pages of an answer sheet) was under 5 seconds, indicating efficient performance for most use cases.

7.2 Ollama 3.1 Text Refinement Module

The Ollama 3.1 Text Refinement Module is used to clean and refine the raw text generated by the Tesseract OCR, addressing common issues such as OCR errors, missing punctuation, and poorly recognized words. This module enhances the readability and accuracy of the extracted text before it is passed to the BERT model for semantic matching.

Results:

- **Text Accuracy Improvement**: The Ollama module successfully improved the raw OCR output by correcting 10-15% of misrecognized characters, fixing sentence structure, and correcting common OCR errors (e.g., replacing “0” with “O” or “I” with “l”).
- **Impact on Grading Accuracy**: With the refined text, the semantic matching module (BERT) achieved a higher degree of accuracy in understanding student responses, reducing errors caused by OCR inaccuracies by 10%.
- **Processing Time**: The text refinement process took an average of 2-3 seconds per answer sheet, ensuring that it did not significantly impact overall performance.

7. 3 BERT Semantic Matching Module

The BERT-based semantic matching module compares the student's refined answer against a predefined answer key, evaluating the content based on meaning rather than exact wording. The key performance metrics for this module are:

- **Accuracy of Semantic Matching**: The percentage of correct evaluations based on semantic similarity between student answers and the correct answer.
- **F1 Score**: A measure of the module's precision and recall in correctly matching the meaning of student answers to the answer key.

Results:

- **Accuracy**: The BERT module demonstrated strong performance in matching semantically similar answers. For answers with different phrasing but similar meaning, the module achieved an accuracy rate of 90%. This was significantly higher than traditional keyword-based matching methods.
- **F1 Score**: The average F1 score for the semantic matching was 0.92, indicating a high level of precision and recall in understanding varied student responses.
- **Handling of Grammatical Variations**: BERT handled grammatical variations well, recognizing correctly phrased answers even when they deviated from the structure of the predefined answer.

7.4 Overall System Performance

The Gradex system, as a whole, demonstrated strong performance in automating the grading process, with a combination of OCR, text refinement, semantic matching, and automated feedback generation. The system achieved:

- **Overall Accuracy**: 92% accuracy in grading, compared to traditional manual grading.
- **Processing Speed**: The system processed each student's answer sheet in under 30 seconds, from OCR to feedback generation, ensuring quick turnaround times.
- **Scalability**: The system was able to handle large volumes of answer sheets, with no significant degradation in performance as the number of submissions increased.