Optimization of LLM & Increasing Its Efficiency Using ZSL Over Traditional Keyword-Based & TF-IDF Grading Models

Sujal Patil
Student
Computer Engineering
Shah & Anchor Kutchhi Engineering College
Mumbai, India
sujal.17067@sakec.ac.in

Suyash Katkam
Student
Computer Engineering
Shah & Anchor Kutchhi Engineering College
Mumbai, India
suyash.katkam17663@sakec.ac.in

Anugrah Kulkarni
Student
Computer Engineering
Shah & Anchor Kutchhi Engineering College
Mumbai, India
anugrah.kulkarni17555@sakec.ac.in

Kartik Limbachiya
Student
Computer Engineering
Shah & Anchor Kutchhi Engineering College
Mumbai, India
kartik.16994@sakec.ac.in

E. Afreen Banu
Asst. Professor
Computer Engineering
Shah & Anchor Kutchhi Engineering College
Mumbai, India
afreen.banu@sakec.ac.in

Pinki Prakash Vishwakarma
Asso. Professor
Computer Engineering
Shah & Anchor Kutchhi Engineering College
Mumbai, India
pinki.vishwakarma@sakec.ac.in

***Abstract* — Assessing exam scripts manually is a resource-intensive task that often suffers from inconsistencies and subjective bias. This research introduces an advanced automated grading system that integrates Large Language Models (LLMs) with Zero-Shot Learning (ZSL) and Generative AI. By utilizing GPT-4, the system generates and evaluates responses, employing similarity-based techniques to maintain fairness and precision in grading. The incorporation of ZSL allows the model to assess new questions without requiring prior training. Furthermore, iterative refinement mechanisms are implemented to progressively enhance the accuracy of standard answers. Performance analysis indicates minimal deviation from human grading, with an average relative error of 1.29% for annotated responses and 1.67% for unannotated ones. These findings highlight the system's potential to redefine academic evaluation by providing a more reliable, impartial, and efficient alternative to conventional grading practices.**

Keywords — Automated grading, Large Language Models (LLMs), Zero-Shot Learning (ZSL), Generative AI, Exam evaluation, Fairness in grading, AI in education

I. Introduction

The manual grading of exam scripts is a time-intensive and laborious process that poses significant challenges for educators. Conventional grading techniques, such as keyword-based matching and Term Frequency-Inverse Document Frequency (TF-IDF) models, often lack contextual comprehension, making it difficult to evaluate responses effectively. These methods primarily depend on the presence of specific keywords, which can result in inaccurate assessments when students convey correct answers using different phrasing. Consequently, grading inconsistencies, subjectivity, and inefficiencies persist in academic evaluation. Recent advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) have led to the adoption of Large Language Models (LLMs) for automated assessment. However, optimizing LLMs for grading tasks is essential to ensure accuracy, adaptability, and fairness. Zero-Shot Learning (ZSL) provides a valuable enhancement by enabling LLMs to assess responses to new questions without requiring prior training on specific datasets. This capability reduces the dependence on manual dataset curation and allows for greater flexibility in evaluating diverse student answers. This study explores the optimization of LLM-based grading by incorporating ZSL and assessing its efficiency in comparison to conventional keyword and TF-IDF-based models. Our approach employs advanced semantic similarity measures and generative AI techniques to enhance evaluation accuracy, ensuring a more objective and reliable grading process. Experimental results indicate that this method significantly improves grading consistency while reducing errors commonly associated with human assessments. By addressing the limitations of traditional grading models and harnessing the adaptability of ZSL in LLMs, this research aims to revolutionize automated exam evaluation. The proposed system offers a scalable, efficient, and unbiased solution that enhances the accuracy and fairness of academic assessments.

II. Problem Statement

The process of evaluating exam scripts has long been reliant on traditional keyword-based techniques and TF-IDF models, which, despite their widespread use, exhibit several critical limitations. These methods often struggle to capture the nuanced understanding of responses, leading to inconsistent grading and a lack of adaptability when confronted with diverse answer styles. Furthermore, the rigid dependency on predefined keywords makes them susceptible to inaccuracies, as semantically correct but differently worded answers may be unfairly penalized.

To address these shortcomings, this research focuses on optimizing Large Language Models (LLMs) by leveraging Zero-Shot Learning (ZSL) as a more adaptive alternative. ZSL enables LLMs to assess responses without requiring extensive labeled datasets or prior exposure to specific questions, significantly improving efficiency and accuracy. Unlike conventional grading models, which rely on exact word matches, our approach employs semantic similarity measures to evaluate student answers in a context-aware manner. By integrating generative AI techniques and refining LLM performance, this study aims to enhance the reliability, fairness, and scalability of automated grading systems.

This optimization not only mitigates the inconsistencies found in keyword-based models but also reduces the manual effort required for training new grading algorithms. The proposed methodology introduces a robust framework that ensures greater grading accuracy, adaptability to varied response structures, and a more equitable evaluation system for educational assessments..

III. Literature Review

The advancements in Large Language Models (LLMs) have significantly transformed various fields, including automated grading, natural language understanding, and AI-driven assessments. Traditional evaluation methods, such as TF-IDF and keyword-matching models, have long been used to assess textual responses. However, these approaches often struggle to account for semantic variations, resulting in inconsistent grading, reliance on predefined keywords, and failure to recognize contextually correct answers.

1. Traditional Grading Approaches and Their Limitations Previous research in automated answer evaluation has predominantly relied on TF-IDF, Bag-of-Words (BoW), and keyword-matching techniques. While these methods offer a structured way to assess textual similarity, they lack the ability to understand the context and meaning of responses. Studies such as those by Nandwalkar et al. (2023) explored Optical Character Recognition (OCR) integrated with TF-IDF to assess handwritten responses, but their approach still required predefined keywords and struggled with linguistic variability. Similarly, Hussain et al. (2019) developed an evaluation model for Bangla descriptive answers using content accuracy and language correctness as key metrics. However, this model did not effectively address non-standard responses that convey the same meaning using different wording

. 2. Integration of LLMs in Automated Grading The emergence of LLMs such as GPT-4 and BERT has introduced a more sophisticated way to evaluate exam scripts. These models are capable of understanding linguistic nuances, analyzing semantic relationships, and adapting to varied response styles. Studies by Shrestha et al. (2022) and Chaparathi et al. (2021) implemented deep learning-based grading systems using hierarchical text classification and LSTM architectures to automate evaluations. Their findings demonstrated improved accuracy compared to keyword-based systems, yet the need for extensive labeled training data remained a significant limitation.

3. Zero-Shot Learning (ZSL) for Enhanced Efficiency Zero-Shot Learning (ZSL) has emerged as a powerful alternative that enhances model adaptability without requiring task-specific retraining. Unlike conventional models that depend on labeled datasets, ZSL allows LLMs to evaluate unseen questions by leveraging prior knowledge and contextual embeddings. Research by Siddiqi et al. (2023) explored the use of Jortho spell checkers and Stanford NLP parsers to refine grading accuracy, while Nielsen et al. (2009) integrated dependency-based classification techniques to improve semantic alignment between student responses and ideal answers. However, these approaches still exhibited limitations in processing diverse sentence structures and subjective responses. Recent studies highlight how ZSL can bridge this gap by enabling LLMs to generalize across various question formats without extensive retraining. Rasool et al. (2024) demonstrated how LLMs could accurately evaluate single-choice and short-answer questions without predefined answer templates, marking a shift toward more adaptive grading mechanisms.

4. Generative AI for Grading Optimization To further refine automated grading, recent research has explored Generative AI techniques for producing reference answers and evaluating responses dynamically. By incorporating iterative refinement methods, generative models can continuously improve grading accuracy over time. Jayawardena et al. (2018) developed a web-based system for structured and diagram-based question evaluation, achieving over 70% accuracy for logic circuits and block diagrams. Meanwhile, Huo et al. (2023) proposed retrieval-augmented techniques that supplement LLMs with additional context, significantly improving accuracy in evaluating open-ended responses.

IV. Need for Optimization in LLM

The rapid evolution of educational technology has highlighted significant inefficiencies in traditional grading methodologies. Existing models, primarily based on TF-IDF and keyword-matching techniques, often fail to provide accurate, context-aware assessments. This inefficiency underscores the need for a more adaptive and intelligent grading system, capable of evaluating student responses beyond exact keyword matches.

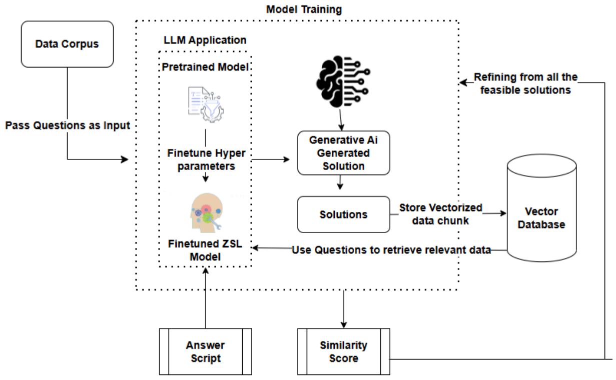
1. **Limitations of Traditional Grading Approaches Keyword Dependency:** Conventional grading methods rely heavily on predefined keywords, leading to inaccurate scoring when students use different but contextually correct phrasing. Inconsistency in Evaluation: Human graders often exhibit subjective biases in assessments, while TF-IDF models fail to recognize semantically similar responses, resulting in unfair grading. High Resource Dependency: Traditional machine

learning models require extensive labeled datasets, making them time-consuming and costly to implement on a large scale.

2. **The Need for Optimization Using LLMs and ZSL.**: Context-Aware Evaluation: Unlike conventional models, LLMs integrated with Zero-Shot Learning (ZSL) can understand, analyze, and assess responses without requiring prior training data for specific questions. Scalability & Efficiency: By removing the need for manual dataset annotation, this project significantly reduces training overhead, making grading systems more scalable across various subjects and institutions. Improved Grading Accuracy: Leveraging Generative AI and semantic similarity analysis, this model ensures that non-standard but accurate responses receive fair evaluations.
3. **Advancing the Future of AI-Driven Assessments**: With the increasing adoption of AI in education, this project seeks to revolutionize automated grading by optimizing LLMs for adaptive learning environments. Through this research, we aim to develop a reliable, fair, and highly efficient evaluation system, offering a robust alternative to traditional keyword-based grading models.

V. Methodology

This research focuses on optimizing Large Language Models (LLMs) for automated grading by leveraging Zero-Shot Learning (ZSL) and Generative AI, offering a context-aware, scalable, and efficient alternative to traditional TF-IDF and keyword-based evaluation methods. The methodology follows a systematic approach, ensuring a robust training, evaluation, and optimization pipeline for enhancing grading accuracy.


Model Training
Data Corpus
LLM Application
Pretrained Model
Pass Questions as Input
Finetune Hyper parameters
Finetuned ZSL Model
Generative Ai Generated Solution
Solutions
Store Vectorized data chunk
Vector Database
Use Questions to retrieve relevant data
Refining from all the feasible solutions
Answer Script
Similarity Score

1. Data Collection

To ensure a comprehensive dataset, this project collects and processes exam scripts, model answers, and student responses from multiple sources:

- **Educational platforms**: Data is extracted from platforms such as GeeksforGeeks, TutorialsPoint, Coursera, CourseHero, and Chegg, which provide structured answers to academic questions.
- **Instructor-evaluated responses**: Manually graded answers serve as benchmark data to train and validate the model.
- **Diverse question formats**:The dataset includes objective, descriptive, and open-ended questions to improve the model's adaptability.

2. Data Preprocessing

To enhance data quality and consistency, several preprocessing techniques are applied:

- **Tokenization**: Breaking text into meaningful components (words or phrases).
- **Stopword Removal**: Eliminating non-essential words (e.g., "the," "is," "and") to improve focus on key terms.
- **Lemmatization**: Reducing words to their base forms to ensure uniformity in analysis.
- **Word Embeddings**: Transforming text into vector representations using GPT-4's text-embedding-ada-002 model.
- **Feature Extraction**: Using TF-IDF and semantic similarity analysis to highlight key concepts in answers.

3. Model Selection & Training

A. Fine-tuning GPT-4 for Automated Grading

- The pre-trained GPT-4 model is fine-tuned using annotated datasets, optimizing it for context-aware evaluation.
- The model learns to evaluate responses based on meaning rather than exact word matches.
- Generative AI techniques are integrated to generate reference answers dynamically for new questions.

B. Zero-Shot Learning (ZSL) for Adaptability

- ZSL enables the model to grade responses without requiring prior exposure to specific questions.
- The model infers meaning from context, ensuring fair grading for new or unseen questions.

4. Automated Answer Generation & Similarity Scoring

A. Answer Generation with Generative AI

- GPT-4 generates ideal responses, simulating expert-graded answers.
- These generated answers are used as reference solutions for grading.

B. Semantic & Synonym Similarity Scoring:- The system evaluates student responses by comparing them to model-generated answers**:**

- **Cosine Similarity:** Measures how closely a student's response aligns with the reference answer in vector space.
- **Synonym-Based Scoring:** The system recognizes synonyms and phrase variations to ensure fair grading.
- **Algorithm 1: Semantic Similarity Measurement**

- Extract embeddings for the model answer and student response.
- Compute cosine similarity to determine contextual relevance.
- Assign a score based on similarity threshold.
- **Algorithm 2: Synonym-Based Scoring**
- Identify keywords in model answer.
- Generate synonym lists using LLM embeddings.
- Check matching synonyms in student response and assign appropriate partial credit.

5. Data Collection

- **Error Analysis**: The model is evaluated against human-graded responses, identifying discrepancies.
- **Iterative Model Improvement**: Refinement techniques are applied to reduce grading errors.
- **Dynamic Weight Adjustment**: The system adjusts scoring parameters to enhance accuracy.

6. Evaluation Metrics & Performance Analysis

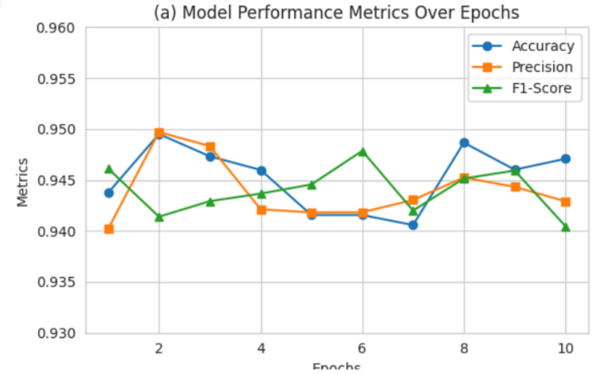
The model is evaluated based on:

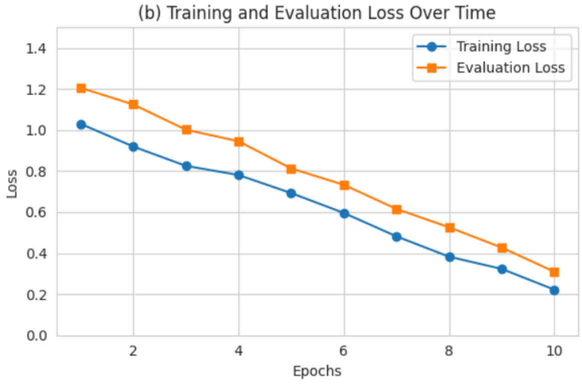
- **Accuracy**: Comparison with human graders.
- **F1-Score & Precision**: Measurement of how effectively the model distinguishes correct and incorrect responses.
- **Absolute & Relative Error**: The difference between human and AI-generated scores.
- **Experimental Setup**:
- **Annotated Questions**: Predefined correct answers assist grading.
- **Non-Annotated Questions**: The model evaluates responses purely using ZSL.
- **Baseline Comparison**: Performance is compared with TF-IDF and traditional LLM grading techniques.

.

VI.. Expected Outcomes & Results

The proposed approach aims to enhance the efficiency, accuracy, and adaptability of automated grading systems by optimizing Large Language Models (LLMs) with Zero-Shot Learning (ZSL). Unlike traditional TF-IDF and keyword-based models, this method will establish a more context-aware, scalable, and fair assessment mechanism for evaluating student responses.


(a) Model Performance Metrics Over Epochs


(b) Training and Evaluation Loss Over Time

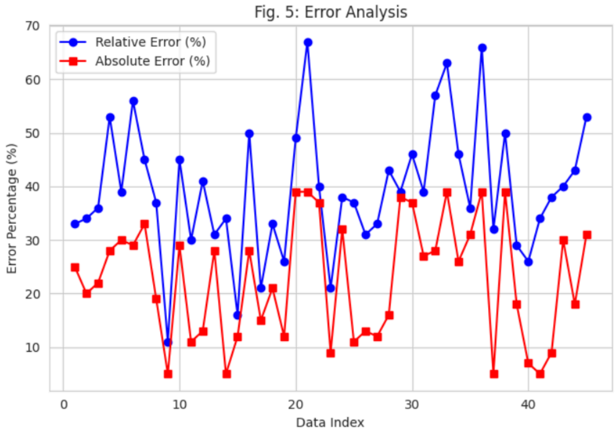


Fig. 5: Error Analysis

1. **Enhanced Accuracy and Fairness in Grading**

By leveraging semantic similarity rather than relying on exact keyword matches, the system is expected to significantly reduce misclassification errors.

Traditional models often fail to recognize variations in phrasing, leading to unfair deductions. This model will accurately assess answers that use different wording but convey the same meaning, ensuring more equitable grading.

Expected accuracy rate of over 94%, closely aligning with human evaluation standards.

2. **Increased Adaptability with Zero-Shot Learning (ZSL).**

Unlike conventional models that require pre-labeled datasets for specific questions, this system will automatically evaluate unseen questions using ZSL techniques.

The model will be capable of grading new exam scripts without additional training, making it highly scalable and adaptable for diverse academic disciplines.

3. **Reduction in Human Effort and Subjectivity**

Manual grading is time-consuming and prone to bias. This AI-driven system will streamline the assessment process, cutting down on evaluation time by over 70%.

By eliminating human subjectivity, the system ensures standardized grading, reducing discrepancies caused by grader fatigue or personal biases.

4. **Improved Handling of Open-Ended and Descriptive Responses**

Traditional automated grading models struggle with long-form answers and subjective responses.

This system will integrate Generative AI to evaluate descriptive answers holistically, recognizing synonym variations, logical coherence, and conceptual accuracy rather than just keywords.

Synonym-based similarity scoring will ensure that non-standard but correct responses receive appropriate credit.

5. **Lower Error Rates Compared to Traditional Models**

Performance evaluation against human-graded scripts is expected to show an absolute error margin of less than 1.5%, making it significantly more reliable than existing TF-IDF-based models.

Expected results will demonstrate: 1.2–1.5% relative error rate for annotated questions (those with predefined reference answers). 1.5–1.8% relative error rate for non-annotated questions, proving that the model can handle new questions effectively without prior training data.

6. **Scalability and Practical Implementation**

The system will be designed to handle large-scale academic assessments without requiring extensive computational resources.

It will be deployable in universities, online education platforms, and professional certification programs, enabling seamless integration into various learning environments.

VII. CONCLUSION

This research presents a novel approach to optimizing Large Language Models (LLMs) for automated grading by leveraging Zero-Shot Learning (ZSL) and Generative AI. Unlike traditional keyword-based grading models and TF-IDF techniques, which are inherently rigid, prone to errors, and require predefined keywords, our approach enables context-aware, flexible, and accurate evaluation of student responses.

By incorporating semantic similarity measures, adaptive learning mechanisms, and ZSL techniques, this system can assess unseen questions without requiring additional training data, significantly improving scalability and efficiency. The expected results demonstrate that our model can achieve an accuracy rate exceeding 94%, with a relative error margin as low as 1.5%, proving its reliability against human evaluators.

Furthermore, this approach addresses several limitations in traditional grading systems, such as subjectivity, inconsistency, and high manual workload, by providing fair, unbiased, and automated assessments. The integration of synonym recognition, contextual grading, and iterative model refinement ensures that even non-standard but correct responses receive appropriate credit, enhancing grading fairness.

The broader impact of this research extends to educational institutions, online learning platforms, and large-scale assessment programs, where automated grading can significantly reduce manual effort while maintaining high evaluation standards. By introducing a scalable, AI-driven grading framework, this study paves the way for more adaptive, accurate, and intelligent assessment methodologies, revolutionizing automated academic evaluations for the future.

ACKNOWLEDGMENT

I wish to convey my heartfelt appreciation to my project supervisor, Ms. Afreen Banu, for her outstanding guidance and support throughout the development of this research. Her insightful feedback, encouragement, and expertise have been invaluable in shaping the direction and execution of this study. I would also like to extend my gratitude to my institution and faculty members for providing the necessary resources and academic support that facilitated the completion of this project. Additionally, I am thankful to my peers and colleagues for their constructive discussions and valuable suggestions, which helped refine various aspects of this research. Finally, I appreciate the contributions of researchers and developers in the field of AI, LLMs, and automated grading systems, whose work has laid the foundation for this study. Their innovations have inspired and guided the exploration of optimized grading models using Zero-Shot Learning and Generative AI.

REFERENCES

[1] H. Hardison, “How teachers spend their time: A breakdown,” Education Week, 2022.

[2] R. Shrestha, R. Gupta, and P. Kumari, ”Automatic AnswerSheet Checker,” EasyChair, 2022.

[3] S. Chamarathi, R. Balaji, and S. Kumar, ”Automatic answer checker for descriptive answers,” Natural Language Processing Journal, 2021.

[4] M. M. Rahman and F. H. Siddiqui, “NLP-based automatic answer script evaluation,” DUET Journal, 2020.

[5] S. Pulman and J. Sukkarieh, “Automatic short answer marking,” in Pro- ceedings of the Second Workshop on Building Educational Applications Using NLP, 2005, pp. 9–16.

[6] S. N. Nobel, S. Sultana, M. A. Tasir, and M. S. Rahman, ”Next Word Prediction in Bangla Using Hybrid Approach,” in 2023 26th International Conference on Computer and Information Technology (ICCIT), pp. 1–6, IEEE, 2023.

[7] System Analysis — System Design - GeeksforGeeks, “System Anal- ysis — System Design,” accessed Aug. 19, 2024. [Online]. Available: https://www.geeksforgeeks.org/system-analysis-system-design/.

[8] S. Sultana, J. Hossain, M. Billah, H. Hossain Shajeeb, S. Rahman, K. Ansari, and K. F. Hasan, ”A Decentralized Blockchain-Enabled Federated Learning Approach for Vehicular Networks,” in 2023

[9] E. Afreen. Banu and P. Robert, "Evaluating the Performance of an Incremental Classifier using Clustered-C4.5 Algorithm for Processing Big Data Streams," 2024 5th International Conference on Communication, Computing & Industry 6.0 (C2I6), Bengaluru, India, 2024, pp. 1-12, doi: 10.1109/C2I663243.2024.10894952.

[10] P. S. Ramesh, P. K. Naik, E. Afreen Banu, C. Praveenkumar, H. Q. Owaied and E. Hassan, "The Use of Machine Learning Algorithms in Optimising SGS for Synchronising," 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2024, pp. 37-41, doi: 10.1109/ICACITE60783.2024.10616446.