



<http://www.diva-portal.org>

This is the published version of a paper published in .

Citation for the original published paper (version of record):

Mughal, N., Imran, A S., Daudpota, S M., Kastrati, Z., Noor, W. (2026)
Exploring potential of large language models for automated essay scoring in education
Discover Artificial Intelligence, 6(1)
<https://doi.org/10.1007/s44163-026-01002-y>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-145293>

RESEARCH

Open Access



Exploring potential of large language models for automated essay scoring in education

Nimra Mughal¹, Ali Shariq Imran², Sher Muhammad Daudpota¹, Zenun Kastrati^{3*} and Waheed Noor⁴

*Correspondence:

Zenun Kastrati

zenun.kastrati@lnu.se

¹Department of Computer Science, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan

²Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway

³Department of Informatics, Linnaeus University, 351 95 Växjö, Sweden

⁴Department of Computer Science, University of Balochistan, Quetta, Balochistan, Pakistan

Abstract

The assessment of open-ended written work is of vital importance to the student learning experience. Conventional essay grading methods heavily depend on expert manual assessment, making them susceptible to errors due to fatigue, bias, and subjectivity. To address this, recent research has introduced AI-based Automated Essay Scoring (AES) systems. While most studies have concentrated on predicting scores, only a few have integrated AES systems with the well-known Large Language Models (LLMs). This study explores the application of LLMs, including GPT and Gemini for AES. The proposed approach was evaluated on two benchmark datasets, namely "Hewlett Foundation: Automated Essay Scoring (ASAP–AES)" and "Learning Agency Lab–Automated Essay Scoring 2.0 (LA–AES)". The proposed method achieved promising results in AES, demonstrating effectiveness on both the benchmark datasets. Statistical analysis revealed that Gemini outperformed GPT, achieving an average Quadratic Weighted Kappa (QWK) score of 0.45 on the ASAP–AES and 0.43 on the LA–AES. To assess the generalizability and objectivity of the proposed approach, real-world data was collected from an O-Level classroom at Sukkur IBA Community College, Pakistan. Multiple human evaluators participated in the study to examine potential biases in human assessment. The findings indicate that LLM-based scoring demonstrates improved objectivity and reduced bias compared to human assessors.

Keywords LLMs, Education, Assessment, Transformers, Writing evaluation, Automated essay scoring

1 Introduction

An essential part of any educational system is the assessment of student's written work for open-ended questions or prompts, for example, essay writing. Such assessments, if correctly done, play a crucial role to quantify a learner's critical thinking ability. However, traditionally essay scoring is done manually by human assessors, which poses a number of challenges and limitations. These include the time-consuming nature of the process, subjectivity, and the possibility of biases such as forbearance or severity bias, individual preferences, and the Halo effect [1]. Inconsistent scoring by human assessors is another significant problem when dealing with a large number of essays. The quality of assessments made by human assessors may be affected by fatigue, lack of sleep, or time



constraints. These factors collectively establish the need for alternative methods that can improve/enhance the essay scoring process that ensure consistency, minimize subjectivity, eliminate biases, and can be scalable. Consequently, Automated Essay Scoring (AES) systems have emerged as potential solutions while offering better and more efficient essay scoring tools [2]. Hence, educators and institutions may utilize these systems to achieve efficient and effective essay scoring, leading to enhanced education quality.

The early AES system tried to solve this problem by using rule-based and machine-learning techniques, mainly with hand-crafted linguistic [3], syntactic [4], semantic [5], and textual features [6, 7]. Word frequency, sentence length, paragraph length, grammatical accuracy, and vocabulary richness are a few of the features of these AES systems. The AES achieved significant results in terms of performance, but as they heavily rely on hand-crafted features, they are difficult to scale and can not adapt themselves to different writing domains [8]. Assessing essays solely on hand-crafted linguistic features may not be sufficient; the inclusion of questions or prompts, along with adherence to rubric guidelines, becomes essential for a comprehensive assessment. Hence, researchers explored revolutionary deep-learning techniques for effective and efficient AES systems in recent years [2].

AES has successfully used recurrent neural networks (RNNs) [9], such as Long Short-Term Memory (LSTM) [10], to capture sequential relationships in essays. Furthermore, a few studies [11–13] examined the effectiveness of transformers for the AES system and provided detailed feedback to students to improve their writing skills. Though Transformers Bidirectional Encoder Representations (BERT) has achieved state-of-the-art performance across various domains of NLP, such as text classification, sentiment analysis, question answering, machine translation, chatbots, and virtual assistants [14], however, in the context of AES, BERT does not provide significant results. Mayfield and Black [15] highlighted that fine-tuning BERT for AES produces equivalent results to those of classical models but at a high computational cost. Further, several researchers suggested that BERT, coupled with handcrafted linguistic and structural features, outperforms other models in terms of performance [13, 16, 17]. However, as discussed earlier, hand-crafted features make it difficult to scale and adapt to different writing domains.

These AES systems suffer from three main limitations. First, no model works on the relevance of content, which means whether a student's response or explanation is relevant to the given prompt and rubrics or not. Second, the generalization, i.e., existing systems trained on specific domains or prompts, face difficulties when applied to essays from different domains. Third, fine-tuning approaches, such as BERT, are computationally expensive with little to no significant performance gain.

Few researchers have attempted to explore the potential of large language models (LLMs) such as GPT for AES in recent works. For example, Song et al. [18] evaluated the performance of open-source LLMs on a dataset of 600 human-scored essays for AES and essay revision tasks. Their findings suggested that LLMs improved the performance of both essay scoring and revision. However, this study had notable limitations, including the use of a limited dataset from a specific domain. Furthermore, the study did not report statistical analyses to support its claims. Similarly, another recent study by Li and Liu [19] evaluated the performance of GPT, BERT, and locally trained LLM for Japanese (Open-Calm large model) for 1400 story-writing scripts by learners with non-native

language speakers. The results revealed that the GPT achieved the highest Quadratic Weighted Kappa (QWK). However, the main limitation of this study is the small data set that limits the generalizability of the model. Furthermore, none of these LLM-based studies have evaluated their approach on benchmark datasets of AES.

A very limited number of studies have explored the potential of LLMs for AES by leveraging their zero-shot learning capabilities, which save time while simultaneously enhancing AES performance. Moreover, very few studies have utilized benchmark datasets in this domain. To the best of our knowledge, only one study by Lee et al. [20] has investigated the zero-shot capabilities of LLMs for AES while employing two benchmark datasets: 1) “Hewlett Foundation: Automated Essay Scoring (ASAP–AES)” and 2) TOEFL11 [21]. Their findings indicate that the small-sized Llama2-13b-chat model outperformed ChatGPT, achieving a QWK of 0.437 on TOEFL11 and 0.355 on the ASAP–AES dataset. Although their results demonstrate significant advancements in AES using LLMs, but exploring open-source models such as Gemini by Google and more recent benchmark datasets can provide more insights for improvement.

Therefore, to overcome the limitations of existing systems, we aim to explore the potential of renowned LLMs, GPT, and Gemini for the AES task. For this, we utilized two benchmark datasets to evaluate the performance of these LLMs, which include: 1) “Hewlett Foundation: Automated Essay Scoring (ASAP–AES)”, a publicly available benchmark dataset related to AES. 2) “Learning Agency Lab–Automated Essay Scoring 2.0 (LA–AES)”¹ that focuses on the evaluation of essays written by students. To evaluate the generalizability of our approach and assess potential biases introduced by human assessors, we collected real-life data from an O-Levels class at Sukkur IBA Community College, Pakistan. Furthermore, we explore and analyze how human assessors may introduce biases in essay scoring and how the quality of the rubric influences the performance of AES systems and the effectiveness of feedback provision.

By leveraging the capabilities of LLMs, we aim to contribute to the advancement of AES systems and facilitate their integration into educational settings. Ultimately, this research aims to improve the efficiency, accuracy, and fairness of essay scoring, empowering educators and students alike.

This study makes four primary contributions to the field of AES:

1. Zero-Shot learning focus: Unlike most prior studies in AES research that rely on fine-tuning or few-shot learning approaches, this study evaluates the zero-shot capabilities of LLMs. This approach is significantly more time-efficient and scalable.
2. Gemini versus GPT comparison: We conduct a rigorous statistical comparison of GPT–3.5-turbo and Google’s Gemini-Pro across two major benchmark datasets, including ASAP–AES and LA–AES, as well as a real-world O-Level classroom dataset collected for the purpose of this study.
3. Impact of rubric quality: This research provides an analysis of how the complexity and quality of rubrics directly influence LLM performance, especially in the case when the rubrics contain subjective expressions, and there is a conflict between the model and a human.

¹<https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2>.

4. Human bias investigation: real-world O-Level classroom data was used in our research to specifically investigate the human-related problems of assessor fatigue and halo effect and compare them to the objectivity of the scoring of the LLM-based scores.

The rest of the paper is organized as follows: Sect. 2 presents the related work in automatic essay scoring. In Sect. 3, we have discussed a detailed methodology describing the LLM prompt. Experiments and results are presented in Sect. 4, whereas Sect. 5 concludes the paper with future directions.

2 Literature review

In this section, we present a comprehensive literature review on automatic essay scoring systems and generative AI for feedback provision. Specifically, we examine the evolution of AES systems based on rule-based approaches [22], statistical models [23], advanced machine learning [24], and deep learning [25] techniques. Furthermore, we investigate the application of transformer-based models [11] in the context of AES. Additionally, we explore the existing research on generative AI models and tools for providing detailed feedback to students [26]. By synthesizing the current literature, we identify the research gap and propose future directions to enhance the accuracy, effectiveness, and efficiency of automated essay scoring and feedback provision systems.

2.1 Evolution of AES

AES is the process of assessing and assigning scores to essays automatically using computational methods. AES system employs algorithms and models to evaluate writings based on preset criteria, and grading guidelines, such as coherence, linguistic correctness, syntax, lexical richness, and semantic relevance [2, 27, 28]. Further, the AES system can also minimize the risks associated with human assessors, such as tiredness, fatigue, and time-constraint [29–31]. As a result, AES systems have gained considerable attention in recent years due to their potential to increase the effectiveness and transparency of essay scoring procedures. In addition, a few studies, such as [32], recently attempted to generate comprehensive feedback for indicating particular areas of strength and weakness in various writing aspects.

Early AES systems used rule-based methods to analyze essays using predefined criteria and grading rubrics. These systems were based on a set of handcrafted rules that captured specific linguistic and structural features [6]. However, such rule-based systems frequently lacked the adaptability needed to manage the complexity and variety of student writing. Hence, statistical models were developed by incorporating linguistic elements into a data-driven methodology such as Latent Semantic Analysis (LSA) [7]. It represents essays and scoring criteria as vectors in a high-dimensional semantic space. Further, it examines the latent semantic connections between essays and the evaluation to compute essay scores, [33]. Similarly, another AES system, E-rater, utilizes a combination of natural language processing (NLP) techniques and statistical models to assess the quality and characteristics of essays [33].

These models enhanced AES's performance, however, these systems were not flexible enough to adapt to the inherent complexity of human language. Further, they also lacked the ability to capture semantic nuances and struggled to generalize across different writing prompts. The field of AES underwent a revolution with the introduction of machine

learning techniques in the late 20th century [34], opening the door for a more robust method.

2.1.1 Machine learning techniques

Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to learn and make predictions or decisions without being explicitly programmed [35, 36]. The usage of feature-based models is one of the well-known machine learning methods in AES. In order to train a regression or classification model to predict essay scores, these algorithms extract a collection of carefully built linguistic [3], syntactic [4], and semantic [5] elements from essays. Word frequency, sentence length, grammatical accuracy, and vocabulary richness are a few examples of these characteristics. For instance, Cohen et al. [37] used features such as sentence length, vocabulary richness, and syntactic complexity to build an Intelligent Essay Assessor (IEA) system. Similarly, Mahana et al. [38] proposed a regression model for AES using the various linguistic and textual features, including Bag of Words (BOW) [39], Parts of Speech (POS) count [40], words count, and Orthography [41].

These machine learning-based methods outperformed prior systems in terms of speed, but they continued to rely significantly on hand-crafted features, which made it difficult to scale and adapt to different writing domains. Hence, researchers explored revolutionary deep-learning techniques for AES systems in recent years.

2.1.2 Deep learning techniques

Deep learning is a type of machine learning that models and solves complex problems such as decision-making, speech and image recognition, natural language processing, and image and audio recognition using vast, numerous hidden layers of neural networks [42]. These networks can automatically learn hierarchical data representations by progressively extracting advanced features from the input raw data. The automated learning of complex patterns and representations from essay writings using deep learning techniques has transformed AES systems [2]. AES has successfully used recurrent neural networks (RNNs) [9], such as the Long Short-Term Memory (LSTM) [10], to capture sequential relationships in essays. These studies focused on predicting the scores closer to human assessors by utilizing state-of-the-art approaches. However, the revolutionary transformers-based models are yet to be explored. Recently, a few studies examined the impact of transformer-based models on the AES system which can also provide detailed feedback to students for improving their writing skills. The subsequent section discusses the architecture of transformers, followed by their implications for AES systems.

2.2 Transformers

Transformers have demonstrated excellent language understanding and language generation capabilities [12], especially Generative Pre-trained Transformers (GPT) [43].

Transformers possess a self-attention mechanism that allows them to recognize word dependencies within a text. Unlike recurrent neural networks (RNNs), which process sequences sequentially, transformers can process words in parallel. The self-attention mechanism can be represented by the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V$$

where Q represents the matrix of query vectors, K represents the matrix of key vectors, and V represents the matrix of value vectors. The dot product between Q and K^T measures the similarity between query-key pairs, which is then scaled by $\sqrt{d_k}$ factor, which is the square root of the dimensionality (d_k) of the query/key vectors. The softmax function is applied to obtain attention weights, which are multiplied element-wise with V to obtain the attention output [44].

The architecture of a transformer consists of an encoder and a decoder, as illustrated in Fig. 1. The encoder maps a sequence of symbol representations (x_1, \dots, x_n) to a sequence of continuous representations ($z=(z_1, \dots, z_n)$). Given z , the decoder then generates a symbol output sequence (y_1, \dots, y_m) one element at a time [44]. In the context of AES, the encoder takes the essay as input and encodes it into a high-dimensional representation. On the other hand, the decoder produces the output, such as essay scoring or providing feedback. The self-attention mechanism of transformers enables both the encoder and the decoder to record global dependencies for the essay and facilitate effective information flow [12].

BERT is a powerful transformer-based model that has revolutionized natural language processing tasks. It includes bi-directional context understanding through the Transformer architecture [45]. BERT has achieved state-of-the-art performance across various

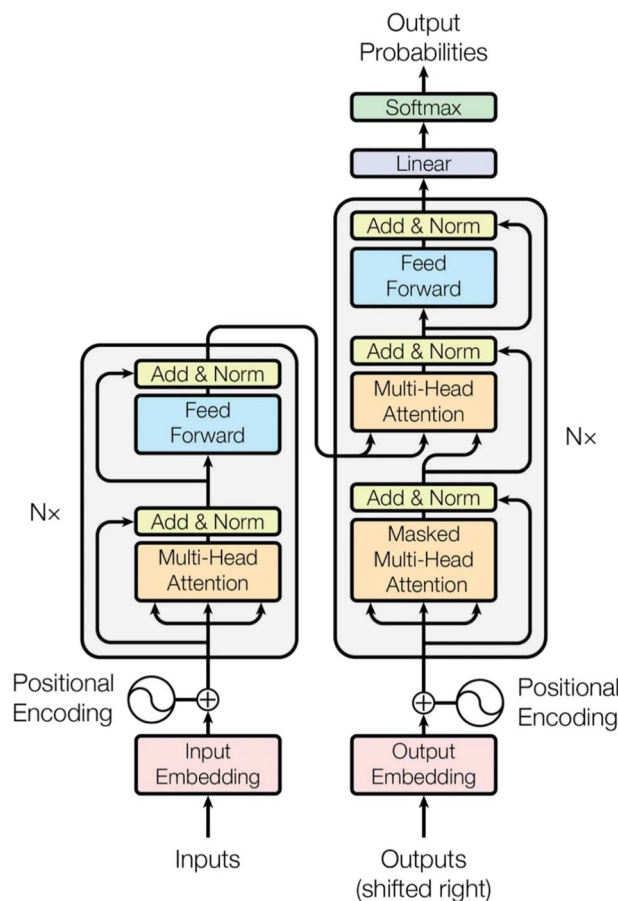


Fig. 1 Transformer architecture [44]

domains of NLP, such as text classification, sentiment analysis, question answering, and machine translation. BERT's contextual understanding of language has led to improved search engines, chatbots, and virtual assistants, enabling more accurate responses and better user experiences [14]. However, in the context of AES, BERT does not achieve significantly better results. Mayfield and Black [15] shows that fine-tuning BERT for AES produced similar results to other models but at a high computational cost. Further, several researchers suggested that BERT, coupled with handcrafted linguistic and structural features, outperforms other models [13, 16, 17]. However, the inclusion of handcrafted features made their scalability and adaptability to different writing domain under question.

Hence, to overcome these limitations, large language models (LLMs) such as GPT can be used to assign the score and provide comprehensive feedback to students to improve their essay write-ups. GPT is a state-of-the-art LLM and probably a watershed moment in the natural language processing (NLP) field [46].

2.2.1 LLMs and AES

Large Language Models (LLMs) are advanced AI models trained on vast amounts of data to understand, generate, and manipulate human language with high fluency and contextual awareness. These models, built on transformer architectures (illustrated in Fig. 1), excel in various NLP tasks such as text generation, translation, summarization, and question answering [47]. Notable LLMs include OpenAI's GPT series [48], Google's Gemini [49], and Meta's LLaMA [50], each contributing to advancements in automated text processing and assessment [51].

Among these, OpenAI² has developed multiple iterations of GPT models that are capable of producing fluent text outputs such as text generation, language translation, and question answering in a human-like manner [46, 52]. Moreover, the latest innovation of ChatGPT³ stunned everyone with its sophisticated features and quickly rose to the top of social media and news outlets. The ChatGPT is uniquely capable of accomplishing incredibly difficult tasks in the education domain, such as writing an article, a story, a poem, or an essay, the ability to provide a summary or expansion of a text, adjusting texts to reflect different perspectives, and even writing and debugging original computer code [53, 54]. While ChatGPT has demonstrated remarkable language generation capabilities and has been applied to various educational tasks, its implementation for AES remains limited and requires further scientific inquiry.

Mizumoto and Eguchi [55] examines the potential of leveraging ChatGPT, for AES in the context of foreign language research, teaching, and learning. The GPT-3 text-davinci-003 model was employed to score a large set of 12,100 essays from the ETS Corpus of Non-Native Written English (TOEFL11) [21] and compared the results to benchmark levels. The study also investigated the influence of linguistic features on AES using GPT. The findings indicate that AES with GPT exhibits a certain level of reliability and accuracy, offering valuable support for human evaluations.

Several recent studies have explored the potential of the LLMs mentioned above for AES. In particular, Song et al. [18] evaluated the performance of open-source LLMs on a dataset of 600 human-scored essays for both AES and essay revision tasks. Their findings

²<https://openai.com/>

³<https://chat.openai.com/>

suggested that LLM improved the performance of both essay scoring and revision. However, the study had notable limitations, including using a relatively small dataset from a specific domain, and it lacked statistical analyses to support its claims rigorously. Similarly, another recent study by Li and Liu [19] investigated the performance of multiple LLMs, including GPT, BERT, and a locally trained Japanese language model (Open-Calm large model), for AES on 1400 story-writing scripts produced by non-native speakers. The results indicated that GPT achieved the highest QWK, outperforming other models. However, the study's main limitation was its small dataset size, which constrained the generalizability of the model's performance. Furthermore, none of these two studies have evaluated their approaches using well-established benchmark datasets for AES, which is crucial for validating model effectiveness.

Latif and Zhai [56] evaluated GPT-4, GPT-3.5, Claude 2, and PaLM 2 on English language learner writing, demonstrating GPT-4's superior intra-rater reliability, though their study was limited to 119 essays from a single placement test and did not examine emerging models like Gemini or rubric quality effects. Similarly, Amin et al. [57] proposed few-shot transformer-based models combining analytical and holistic scoring approaches, though their work focused primarily on few-shot rather than zero-shot learning scenarios.

Despite these recent advances, several research gaps remain: a very limited number of studies have explored the potential of LLMs for AES, particularly their zero-shot learning capabilities. Most prior AES models relied on traditional fine-tuning of LLMs and the few-shot learning techniques. Moreover, few studies have rigorously evaluated LLMs on benchmark datasets commonly used for AES research. To the best of our knowledge, only one study by Lee et al. [20] has investigated the zero-shot capabilities of LLMs for AES while employing two benchmark datasets: (1) "Hewlett Foundation: Automated Essay Scoring (ASAP-AES)" and (2) TOEFL11 [21]. Their findings indicate that the small-sized Llama2-13b-chat model outperformed ChatGPT, achieving a QWK of 0.437 on TOEFL11 and 0.355 on the ASAP-AES dataset. These results highlight the potential of smaller open-source models in AES tasks, suggesting that fine-tuned or domain-adapted models may outperform general-purpose proprietary models. However, their study was limited to a small selection of models and datasets. Exploring additional open-source models, such as Gemini by Google, along with more recent and diverse benchmark datasets, could provide further insights and improvements in AES performance.

Hence, to address the limitations of existing AES systems, we aim to explore the potential of renowned Large Language Models (LLMs), specifically GPT and Gemini, for the AES task, with a particular focus on their zero-shot capabilities. Additionally, this study investigates how human assessors may introduce biases in essay scoring by utilizing real-life data from an O-Levels class at Sukkur IBA Community College, Pakistan. Furthermore, we examine the impact of scoring rubric quality on the performance of AES systems and the effectiveness of feedback provision.

3 Methodology

In this section, we discuss the methodology employed to evaluate the potential of LLMs, specifically, GPT-3.5-turbo and Gemini-pro for Automated Essay Scoring and feedback provision to students. Four main steps are carried out on the collected datasets, as depicted in Fig. 2.

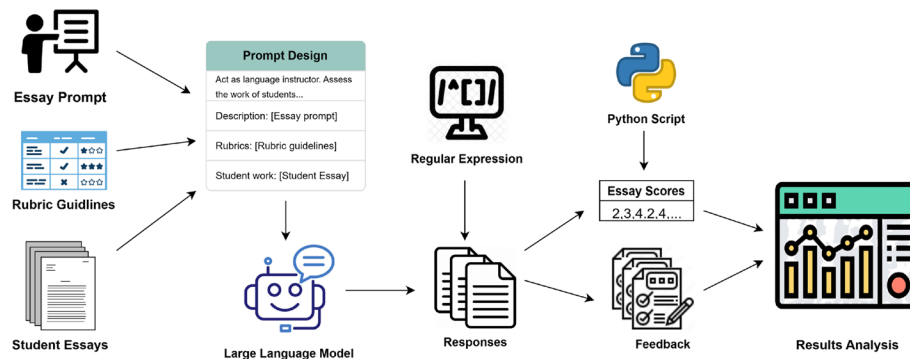


Fig. 2 Methodology for AES evaluation

Assessment checker

The screenshot shows the 'Assessment checker' interface. On the left, there are three input fields: 'Enter Description:', 'Enter Rubrics:', and 'Enter Text for Assessment:'. The 'Enter Description:' field contains the text: 'nature, and interacting with family and friends. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.' The 'Enter Rubrics:' field contains the text: 'Rubric Guidelines Score Point 1: An undeveloped response that may take a position but offers no more than very minimal support. Typical elements:'. The 'Enter Text for Assessment:' field contains the text: 'because it can breakdown. Dats one reason why computers arent so good. But as j said computers are very good they might be a little pesky but ones you get the hand of them everything gonna be alright and remember you can do almost anything with a computer.' Below these fields are 'Generate' and 'Print' buttons. On the right, the 'Response by GPT' section shows the output: 'Score: 2', 'Comments:' (listing four points about organization, elaboration, position, and language), and 'Highlighted mistakes:' (listing five grammatical errors).

Fig. 3 Interface of assessment checker

First, we design a prompt for LLMs, including the essay prompt/description, rubrics, and essays composed by students. Second, we made API calls in the loop to get the essay score and feedback for all the essays. Third, we process the response files using regular expressions to extract the essay score and auto-generated feedback. Finally, we present results after analyzing the scores and auto-generated feedback. The subsequent sections present a detailed description of the methodology from data acquisition to results analysis. In addition, we also designed a Graphical User Interface (GUI) for the AES system and auto-generated feedback using Flask,⁴ which is displayed in Fig. 3. The user can feed rubrics and assignment descriptions as per the choice. Hence, our designed approach may be evaluated for various kinds of rubrics and essays.

3.1 Dataset acquisition

In this study, we used two standard datasets for essay scoring: *Hewlett Foundation: Automated Essay Scoring dataset (ASAP–AES)*, a state-of-the-art dataset, and *Learning Agency Lab–Automated Essay Scoring 2.0 (LA–AES)*. ASAP–AES dataset provides a vast collection of essays written by students across various academic levels and disciplines. Each essay is accompanied by a score assigned by expert human assessors,

⁴<https://flask.palletsprojects.com/en/2.3.x/>

ensuring the availability of reliable and accurate ground truth. In addition to these benchmark datasets, we also evaluated the performance of data collected from a real-life classroom scenario. A detailed description of these three datasets is provided in subsequent sections.

3.1.1 ASAP–AES dataset

The dataset covers eight diverse topics for essays, facilitating the examination of the generalizability and robustness of the developed models. This dataset also provides the essay descriptions and rubrics for each set.

We selected six essay sets from this dataset, as these six sets had a score range 0–6, whereas the remaining two sets had a huge variation in score range i.e. 0–30. The dataset description of six chosen essay sets is summarized in Table 1. Each essay was assessed by two human assessors and in the case of any disagreement, the criteria were defined to get the resolved score. In some of the sets, the first assessors' score was given preference, whereas in some sets another domain expert was involved to assign the resolved score as per the rubric. All of the sets had identical patterns to get the resolved score, except set-1, which had a score range of 1–6 for each human assessor. For this set, the resolved score was the summation of two human assessors' scores in case of identical or adjacent scores. Whereas, a third expert was involved in case of disagreements. Therefore, the resolved score was out of the rubric range, of 1–6, for most of the essays. In addition, Essay set-2 has two different scores, the domain 1 score represents the writing implication score, and domain 2 represents the language convention score. Combining these two scores may not be appropriate as both domains had different ranges, i.e. 1–6 and 1–4, respectively. Further, these essays lie in the two main categories i.e. persuasive, and source-dependent essays:

1. *Persuasive Essay* A persuasive essay attempts to persuade the reader to embrace or agree with a particular point of view or to perform a specified action [58]. The writer uses arguments, evidence, and logical reasoning to defend their position and persuade the reader that it is correct. Persuasive essays frequently employ persuasive strategies such as emotional appeals, logical appeals, and credibility appeals to strengthen the argument [28]. The essay is organized in a logical manner, with an introduction, body paragraphs offering arguments and evidence, and a strong conclusion that restates the important points and confirms the writer's perspective. Essay set-1, and set-2 consist of persuasive essays; set-1 is written by Grade-8 students on the topic of "The Impact of Computers on Society" and set-2 is written by Grade-10 students on the topic "Censorship in the libraries".

Table 1 Detest description, Dom1 represents writing implications and Dom2 represents language convention

	Essay type	Count	Mean Score	Rubric Range
Essay set-1	Persuasive	1783	8.53	1–6
Essay set-2	Persuasive	1800	Dom1: 3.42 Dom2: 3.33	Dom1:1-6 Dom2:1-4
Essay set-3	SDR	1726	1.85	0–3
Essay set-4	SDR	1771	1.43	0–3
Essay set-5	SRD	1805	2.41	0–4
Essay set-6	SDR	1800	2.72	0–4

2. *Source Dependent Response* Source Dependent Response (SDR) essays are frequently employed in standardized testing, particularly in response to reading comprehension passages. In an SDR essay, test participants are given a text or texts to examine and react to certain questions or prompts depending on the information presented in those texts. The main feature of an SDR essay is that it is highly dependent on the offered sources or texts to develop a response. To support their responses, test takers are asked to demonstrate their grasp of the sources, accurately describe essential themes, and incorporate significant information from the texts. It is critical to attentively evaluate the offered texts, identify the main concepts, and assess any supporting evidence or examples presented in an SDR essay. The response should be focused, well-structured, and address the prompts or questions directly. Also, it is critical to effectively explain ideas and provide clear explanations or arguments based on the material presented in the sources in an SDR essay. Except for set-1 and set-2, all the remaining sets comprise SDR essays on various topics, including stories, fiction, and articles.

3.1.2 LA-AES

This dataset is taken from the Kaggle competition “Learning Agency Lab–Automated Essay Scoring 2” <https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2> and focuses on the evaluation of essays written by students. It contains essays written in response to a variety of prompts, with scores provided by human assessors. The dataset aims to support the development of automated scoring systems that can predict the quality of an essay based on specific criteria. It includes various features such as the essay text, scores assigned for holistic quality, and sometimes more granular aspects like grammar, content relevance, and structure. This dataset is intended for researchers and data scientists working on natural language processing (NLP) tasks, machine learning models, and automated grading systems. It serves as a benchmark for assessing how well the automated methods can replicate human judgment in evaluating the quality of student writing. The scoring range for this dataset is from 1 to 4, as depicted in Table 2. Additionally, the table presents the number of essays corresponding to each score. We have selected 30% of the essays from each score category using stratified random sampling. This approach was chosen to ensure representative distribution across all score categories while reducing the total volume to a manageable size for API-based evaluation, as mentioned in Table 2.

3.1.3 Real-life dataset

Apart from the benchmark dataset used in our experiments, we also evaluated our proposed model for the O-Levels English class (subject code: 1123) of the Sukkur IBA Community College, Pakistan. The evaluation encompassed a diverse range of essays covering various topics and writing styles. Specifically, we collected 21 essays written

Table 2 Number of essays per score in the original and selected datasets

Score	Original	Selected
1	1252	375
2	4723	1416
3	6280	1884
4	3926	1177
Total	17,181	5143

by students from Google Classroom from different writing domains, including informal writing, report writing, narrative writing, and descriptive writing. Two domain experts, including the subject teacher (Human Assessor1, HA₁) and an educator from another class (Human Assessor2, HA₂), participated in the evaluation process to provide a comprehensive perspective on the quality of the proposed approach.

3.2 LLMs

In this study, we use two well-known LLMs, namely GPT and Gemini. A detailed description of both models is mentioned in subsequent sections.

3.2.1 GPT model

In this study, we use GPT-3.5-turbo, which was released on 23 March 2023 by the OpenAI community. This is the well-known and renowned model of GPT that allows access through API keys. By utilizing this model, we predict the score for the essay and also generate detailed feedback to further enhance the effectiveness of GPT models for assessment. We automatically score each essay with the GPT-3.5-turbo API from OpenAI. API requests are sent with parameters and prompt design using Python 3.9. The cost for using OpenAI's API is \$0.002 per 1000 tokens.

3.2.2 Gemini model

Gemini is an advanced generative AI model developed by Google AI. In this study, we used an API version of the model, Gemini-Pro.⁵ This versatile LLM excels in complex reasoning, instruction following, code generation, and multi-turn conversations. It builds upon the foundational research of the Gemini family of models, demonstrating advanced capabilities in various text-based tasks.

3.3 Prompt design

While designing the prompt for LLM models, we closely observed the practices that human assessors follow to assign the score. The same prompt was used for Gemini-Pro and GPT-3.5-turbo. For the ASAP-AES dataset, human assessors are provided with a detailed essay description/prompt that was given to students for composing an essay and clear rubric guidelines to evaluate various components of writing. Hence, we precisely designed the prompt ASAP-AES dataset by including the rubrics and description, as shown in Fig. 4. Further, for the LA-AES dataset, the rubrics were provided to assess the scores regardless of the prompt/description for writing the topics. Hence, we designed the clear and precise prompt for LA-AES with the detailed instructions, as illustrated in Fig. 5.

3.4 Data privacy and ethics

In the research on educational AI, the primary concern is the responsible use of student data. In the publicly available benchmark datasets (ASAP-AES and LA-AES), all the essays were already anonymized and made available for research use. In our case of O-Level real-world data that was obtained at Sukkur IBA Community College in Pakistan, we also applied rigorous anonymization measures. Before processing the

⁵<https://blog.google/technology/ai/google-gemini-ai/>

```
Act as an experienced language instructor and assess the work of students. I will provide you the assignment description/task given to students.
Then, I will provide the students' work and the rubrics.
Your task will be to assign points and comments against each rubric first.
Then, highlight words, sentences, or phrases where the student made mistakes or violated rubrics.
Assignment description:
{description};
Rubrics:
{rubrics};
Student Work:
{text}
```

Fig. 4 Prompt used for ASAP–AES data

```
Act as an experienced language instructor and assess the work of students.
I will provide the student's essay and the rubrics.
Your task is to assign a score and provide detailed comments for each rubric.
Highlight words, sentences, or phrases where the student made mistakes or did not meet the rubric criteria.
Rubrics:
{rubrics}
Student Work:
{text}
Instructions:
1. Assign a score for each rubric criterion based on the student's essay.
2. Provide detailed comments explaining the score for each criterion.
3. Highlight specific mistakes or areas where the essay did not meet the rubric criteria.
4. Offer constructive feedback for improvement where applicable.
```

Fig. 5 Prompt used for LA–AES dataset

responses, all the responses of students were provided with specific identifiers (e.g., e001 to e021) to remove any personally identifiable information (PII). The API calls to OpenAI and Google did not transmit any student names, identification numbers, or any other personal data that could be used to identify a student. This research follows the ethical principles of research and guarantees that the privacy of the students is upheld during the assessment.

3.5 Experimental setup

A Python script is written to retrieve the assignment description, assignment rubrics, and student work from CSV and Word files, as well as to assign scores to each essay. Python 3.9 and OpenAI's GPT-3.5-turbo, and Google's Gemini-pro models are used to generate the results. Both models were configured with a temperature of 0.1 to ensure that LLM results are consistent and deterministic, which minimizes randomness in the generated outputs. API calls are made in the loop to retrieve the scores and feedback for all essays. Further, regular expressions are used to extract the scores and feedback. The confusion matrix is used to compare the score assigned by the human assessor and the score predicted by LLMs. By comparing the predicted scores to the actual scores, we can identify specific patterns of misclassification, such as cases where the model consistently overestimates or underestimates the scores. Further, Quadratic Weighted Kappa (QWk)

scores are computed for statistical analysis and a comprehensive analysis between humans and LLMs is presented.

4 Results and discussion

This section presents a detailed analysis of the results produced by the GPT-3.5-turbo and Gemini models for AES. As mentioned in the description of the ASAP-AES dataset, six different sets of essays were selected for experimental analysis. Of these selected essays, a few of the essays could not be scored due to the token size restriction when sending prompt data to LLMs using the API. Table 3 shows the total number of essays and the number of essays that the GPT and Gemini models could assess for each dataset. It can be seen that very few essays could not be processed overall. However, it is also noticeable that essay set-2 has two different domains, i.e., writing implications and language conventions. More than 200 essays could not be scored for domain2 (language convention) due to API token limitations, as the rubric guidelines for domain2 contains more words as compared to domain1. Further, it can be also observed from Table 1 that selected essays predominantly fall into two primary categories of essays, namely Persuasive and Source Dependent. Hence, the subsequent sections provide a detailed analysis specifically focusing on these two essay types.

4.1 GPT results for ASAP-AES dataset

Essay set-1, and set-2 consist of persuasive essays on the topics of “Effects of Computers on People’s Lives” and “Censorship in Libraries”, respectively. In terms of the rubric range, set-1 contains values between 2 and 12. However, set-2 contains two different domains, each with its own unique range of scores. Specifically, the “Writing Implications” domain ranges from 1 to 6, while the “Language Conventions” domain ranges from 1 to 4, as indicated in Table 1. The performance of the GPT model exhibits variations when applied to two sets, indicating its sensitivity to the characteristics and content of the input data. It shows significant results and noticeable results for set-2. However, in the case of set-1 GPT model was unable to produce results closer to human assessors.

In contrast, Essay set-3, set-4, set-5, and set-6 consist of SDR essays based on various stories that were provided as input to the students. All these sets were written by grade 10 students except set-5, which was written by grade 8 students. In terms of rubric range, set-3 and set-4 contain values between 0 and 3. However, set-5 and set-6 contain values between 0 and 4, as indicated in Table 1. The results produced by the GPT model were closer to human assessors for most of the essays with a maximum variation of point 1. A detailed analysis of these findings is discussed below.

Table 3 No. of ASAP-AES and LA-AES Essays assessed by GPT and Gemini

Dataset	EssaySet	Original	Scored by GPT	Scored by Gemini
ASAP-AES	Set-1	1783	1536	1685
	Set-2-domain1	1800	1726	1525
	Set-2-domain2	1800	1575	1524
	Set-3	1726	1691	1544
	Set-4	1771	1733	1558
	Set-5	1805	1804	1277
	Set-6	1800	1775	1320
LA-AES	–	5134	4846	5074

4.1.1 Essay set-1

The essays were written on the topic of “Effects of Computers on People’s Lives” by students in grade 8. As mentioned in the data acquisition section, the resolved score exceeded the specified range of rubrics for most of the cases due to the summation of scores assigned by two human assessors. Hence, to facilitate a comparison of the results produced by GPT, we calculated the average score of the two human assessors. Score predicted by GPT was in the range of 1–5 for the majority of the essays. However, it is worth noting that 20 outlier results were produced where the predicted score exceeded 6, indicating a huge deviation from the expected range. These outlier results indicate the evaluation of four writing elements: elaboration, organization, fluency, and audience awareness, which were mentioned in the rubrics. GPT assigned the score to all four elements individually in the range 1–6. Therefore, when these individual values are summed, they may exceed 6 or the expected range, as demonstrated in Fig. 6. Hence, to ensure compatibility with the confusion matrix and improve visual clarity, the outlier scores for the 20 instances were divided by 4. This adjustment ensures a consistent representation and facilitates a better interpretation of the data.

The comparison between the scores of the human assessors and GPT is shown in Fig. 8a. The results indicate that GPT could not produce promising results for a majority of the essays. Only a few essays were assigned score closer to human assessors, whereas in most cases, there was a variation of 2 or 3 score points. For instance, GPT assigned a score of 2 to 706 essays, whereas the human assessor had assigned 4, which indicates a variation of 2 points. Similarly, 251 essays were assigned a score of 2 by GPT, whereas the human assessor’s score was 5, which indicates the variation of point 3. The variations between the scores of the human assessors and GPT are illustrated in Fig. 8b. Only 4.9% of the essays had the exact same score assigned by GPT and human assessor, while 21% of the essays had a variation of 1 point. In contrast, the majority, 52.3% of the essays, had a variation of two points. Consequently, it was concluded that the GPT model could not

Essay_id = 907, Human ₁ Score=5, Human ₂ Score=4, GPT-3.5-turbo Score=14
<u>The response generated by GPT Model:</u>
<p>Rubric Assessment: Content: 4/6 Organization: 3/6 Fluency: 3/6 Audience Awareness: 4/6 Overall Score: 14/24</p> <p>Comments: - The student has taken a position on the topic but has provided minimal support for their argument. - The organization of the letter is weak and lacks clear transitions between ideas. - The fluency of the writing is also weak and the language used is simplistic. - The student shows some awareness of the audience but could have done more to persuade them.</p> <p>Mistakes/ Violations: - "enformed" should be "informed" - "buisness" should be "business" - "relize" should be "realize" - "imfortant" should be "important" - There are several instances of run-on sentences that could</p>

Fig. 6 Example for outlier data

produce promising results for set-1, which may indicate the poor rubric quality to assess the essays.

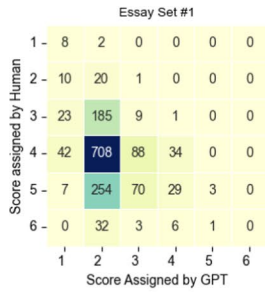
We analyzed the essay descriptions and rubrics to understand the reasons for such significant score variations and outlier scores that was beyond the actual score of the human assessors. It was observed that the rubrics for this essay set were complex and less quantifiable, and to this extent, GPT could not produce close results to those of human assessors. For instance, it was mentioned in the rubrics that four elements of writing (elaboration, organization, fluency, and audience awareness) should be considered. However, it was also mentioned later that individual responses may be stronger in one aspect and weaker in another, and that a score point cannot perfectly describe every response within a score-point category. As a result, human assessors may considered one of the components while assigning the score. Meanwhile, GPT considered and mentioned all the components while assigning the score and providing detailed feedback to students by highlighting mistakes. Figure 7 illustrates an example of LLM assessment, highlighting the inclusion of four key elements. The figure demonstrates how GPT provided suggestions for improvement, specifically in elaboration, organization, and audience awareness. Further, the presence of outlier data also indicates the inclusion of these four elements, as mentioned in the above paragraph.

Essay_id = 1758, Human ₁ Score=5, Human ₂ Score=5, GPT-3.5-turbo Score=2
<u>The response generated by GPT Model:</u>
<p>Rubric Assessment: Score Point: 2 Overall Statement: An under-developed response that may or may not take a position. Typical Elements: - Contains only general reasons with unelaborated and/or list-like details. - Shows little or no evidence of organization. - May be awkward and confused or simplistic. - May show little awareness of audience.</p> <p>Comments: - The response lacks a clear introduction and conclusion. - The reasons provided are not well-elaborated and lack specific details. - The organization is poor, and the response is confusing and awkward to read. - The response shows little awareness of the audience.</p> <p>Mistakes: - "imnarred" should be "impaired." - "kidnaping" should be "kidnapping." - "inapropiate" should be "inappropriate." - "obeise" should be "obese." - "exersice" should be "exercise." - "Preety" should be "Pretty." - "becomeing" should be "becoming." - "cataraks" should be "cataracts." - "althritis" should be "arthritis." - "fourty" should be "forty."</p> <p>Overall, the response needs significant improvement in terms of organization, elaboration, and awareness of the audience. The student should work on providing specific details to support their reasons and organizing their ideas in a clear and logical manner. Additionally, they should proofread their work to correct spelling and grammar mistakes.</p>

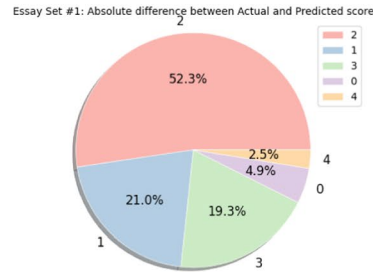
Fig. 7 GPT assessment example for essay set-1

4.1.2 Essay set-2

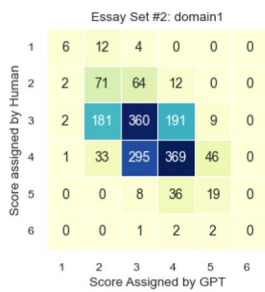
The essays were written on the topic of “Censorship in Libraries” by students of grade-10. Essays were assessed against two different components, including writing implications and language conventions. The comparison between the scores of the human assessors and GPT is shown in Fig. 8c for the *writing implication* domain, which focuses on evaluating how well essays convey their intended meaning. . The results indicate that



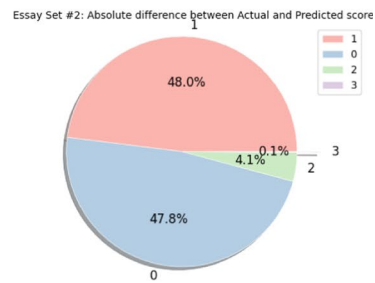
(a) CM - Set-1



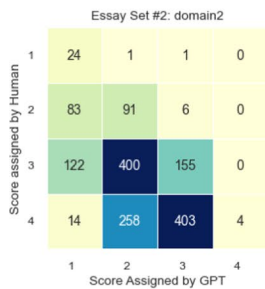
(b) Pie Chart - Set-1



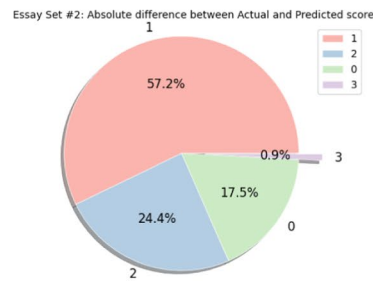
(c) CM - Set-2 Domain-1



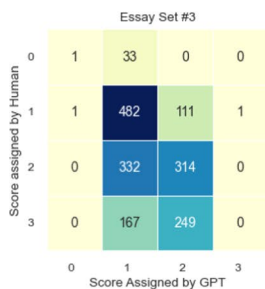
(d) Pie Chart - Set-2 Domain-1



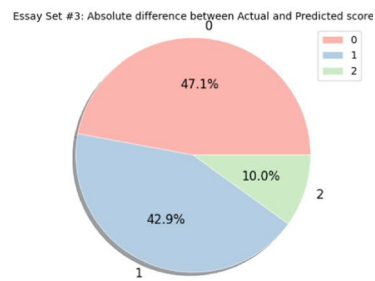
(e) CM - Set-2 Domain-2



(f) Pie Chart - Set-2 Domain-2



(g) CM - Set-3



(h) Pie Chart - Set-3

Fig. 8 GPT results for ASAP–AES dataset dataset (EssaySet 1, 2, and 3)

GPT produced promising results and predicted scores were closer to human assessors, with a maximum variation of point 1 for majority of the essays. For example, the GPT and the human assessors both assigned a score of 3 to 360 essays, indicating a variation of 0. Furthermore, 181 essays were assigned a score of 2, and 191 essays were assigned a score of 4 by GPT, where the human assessors' score was 3, indicating the variation of point 1. In addition, the results for language conventions are presented in Fig. 8e. A similar pattern can be observed for language convention also with a maximum variation of point 1 for most of the essays. However, a significant number of essays showed the variation of point 2, which is a noticeable variation in the score range of 0–4. As can be observed from Fig. 8e, 122 essays were assigned the score of 1 by GPT, where the human assessor's score was 3, indicating the variations of point 2. Similarly, GPT assigned a score of 2 to 258 essays, whereas a human had assigned a score of 4.

The general variations between the scores of the human assessors and the GPT are illustrated in Fig. 8d for the writing implication tasks. Results indicate that Writing implications scores assigned by GPT were closer to human assessors for the majority of the essays with a maximum variation of point 1. Only 4.1% and 0.1% essays exhibit a variation of points 2 and 3, respectively. In contrast, the scores for language convention showed a variation of points 2 and 3 for 24.4% and 0.9% respectively as illustrated in Fig. 8f.

The thorough analysis was performed in order to investigate the possible reasons for better performance for writing implications as compared to language conventions. It was observed that the rubrics for Writing implications were very detailed and quantifiable. Each four components of Writing was explicitly mentioned in the rubrics, and clear instructions were provided to assess these four components of writing. It includes organization, ideas and content, writing style, and voice of the essay. In contrast, the rubrics for language conventions were generic, including capitalization, punctuation, spelling, grammar, paragraphing, and sentence structure. These rubrics employed four keywords- 'minimal', 'fair', 'good', and 'Superior' - to assign scores ranging from 1 (minimum) to 4 (maximum). For instance, it was mentioned in the rubrics that a score of 4 should be assigned if the writing sample demonstrates a superior command of the aforementioned language skills. Similarly, if a writing sample demonstrates minimal command than 1 score should be assigned. Such terms are very subjective even different human can extract different meanings from them. Human assessors may ignore or overlook some of the errors as these four terms are very subjective and may indicate different meanings in different scenarios.

4.1.3 Essay set-3

This set consists of SDR essays that were written in response to the story "ROUGH ROAD AHEAD: Do Not Exceed Posted Speed Limit". The students were supposed to write a response to explain how the characteristics of the environment affect the cyclist. The results presented in Fig. 8g indicate that the GPT model achieved promising results for these SDR essays with a maximum variation of point 1 for most of the essays. For example, the GPT and the human assessors both assigned a score of 1 to 482 essays, indicating a variation of 0. Furthermore, 332 essays were assigned a score of 1 by GPT, where the human assessors' score was 2, indicating the variation of point 1.

In general, 47.1% and 42.9% essays exhibit variations of 0 and 1, respectively, as illustrated in Fig. 8h. On the contrary, 10% of the essays exhibit variations of point 2. More precisely, a total of 169 essays exhibit a variation above 1. Of which 167 essays were underestimated by the GPT model by assigning a score of 1 to the essays that were scored 3 by human assessors as indicated in Fig. 8g. Therefore, we conducted a thorough analysis on these 167 essays to investigate the potential factors contributing to the observed variation.

It was observed that GPT prioritized grammar, punctuation, spelling, coherence, and clarity as key factors. A total of 133 of the 167 essays exhibited errors in these areas. On the contrary, the rubrics used in the evaluation process lacked specific guidelines for these writing components, which could lead to oversight or dismissal of such errors by human assessors. In contrast, for the remaining 34 essays, GPT emphasized task completion and the details added by students to justify the presented ideas. An example is shown in Fig. 10. It can be observed that the response submitted by the student lacks details. The average length of the essays for this set is 150 words, whereas this example contains only 50 words. However, human assessors may unintentionally overlook such errors due to various factors, such as halo effects, leniency or severity biases, and personal preferences.

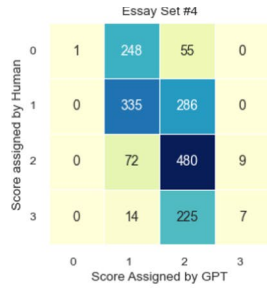
4.1.4 Essay set-4

This set consists of the SDR essays that were written by grade-10 students in response to the story “Winter Hibiscus by Minfong Ho”. Students were supposed to respond to the ending of that story. The GPT assigned scores were closer to human assessors for most of the essays with a maximum variation of point 1. For instance, both GPT and human assigned a score of 1 to 335 essays, indicating a variation of 0. Furthermore, 286 essays were assigned a score of 2 by GPT, which indicates the variation of point 1 as illustrated in Fig. 9a. Variation of the above 2 was found for very few essays. More precisely 70 essays exhibit a variation of 2, which is 4% of this set as depicted in Fig. 9b.

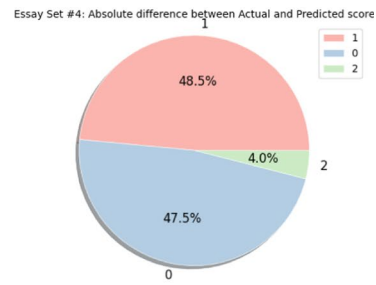
Although few essays had variation exceeding 1, we still examined these essays and corresponding results thoroughly to understand the underlying cause of the error. Of these essays, 14 were underestimated by GPT and the remaining were overestimated. The reasons for underestimation were the same as of set-3, GPT put strict compliance on grammar, punctuation, spelling, coherence, and clarity, and details were added to support the answer. Whereas, GPT assigned a score of 2 to those essays where few details were added, which is not logically incorrect. However, on closer examination of such results, it was revealed that GPT put more emphasize on grammar and other writing components as compared to the details and evidence provided by students to support the point. The example presented in Fig. 11 indicates that GPT highlights the need to add more depth and evidence. Whereas, due to no grammar and spelling errors, it has assigned a partial score to the response. In contrast, human assessors had assigned a 0 score to this essay as they may have emphasized more on the detail and evidence added by students.

4.1.5 Essay set-5

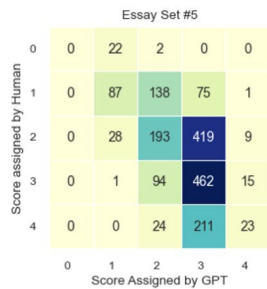
This set consists of SDR essays that were written in response to a memoir “from Home: The Blueprints of Our Lives”. The students were supposed to describe the mood created by the author by supporting their responses with relevant and specific information from



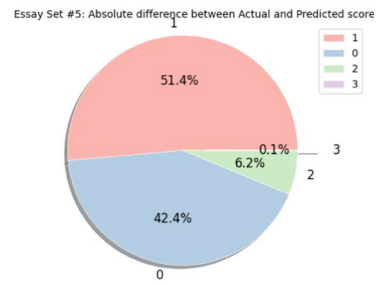
(a) CM - Set-4



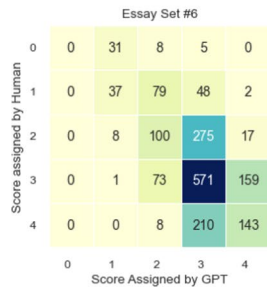
(b) Pie Chart - Set-4



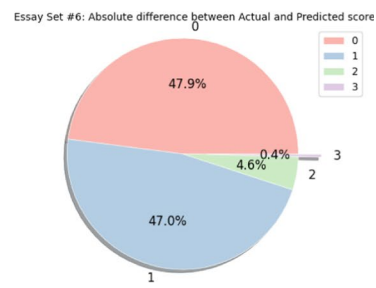
(c) CM - Set-5



(d) Pie Chart - Set-5



(e) CM - Set-6



(f) Pie Chart - Set-6

Fig. 9 GPT results for ASAP–AES dataset (EssaySet 4, 5, and 6)

the memoir. GPT model produced great results with this set also, with a maximum variation of 1 for most of the essays as depicted in Fig. 9c. However, it can be observed that 75 essays were overestimated by GPT by assigning a score of 3 whereas human assessors had assigned a score 1. This indicates a variation of 2. Overall, variation for this set is indicated in Fig. 9d, which shows that a total of 6.3% had variation above 1. Hence, those 75 essays and the corresponding results were closely examined, and it was found that the reasons were the same as those of set-4. GPT puts more emphasize on the clarity and grammar structure however the details provided by students were given less attention.

4.1.6 Essay set-6

This set consists of SDR essays that were written in response to a report “The Mooring Mast by Marcia Amidon Lusted”. The students were supposed to describe the challenges, the builders of the Empire State Building faced. A similar pattern was observed for the score variation as of set-4, and set-5. Overall, 4.6% of the essays showed a variation of 2, while only 0.4% of the essays showed a variation of 3 (Fig. 9f), more precisely 82, and 7 essays respectively (depicted in Fig. 9e). Of these 89 essays, only 16 essays were

Essay_id = 7513, Human₁ Score=3, Human₂ Score=3, Resolved =3 , GPT-3.5-turbo Score=2
<u>Student Work</u>
The cyclist became very tired and the area he was in did not help at all. he was in the desert in june very hot day and every town he had come to was abandoned and had no water. It was also very hilly which made the work even harder.
<u>GPT Response</u>
<p>Rubric Assessment: Score: 1</p> <p>The response shows a minimal understanding of the text. The student has addressed the demands of the question but has not developed all parts equally. The response lacks information or explanation to support an understanding of the text in relation to the question.</p> <p>Feedback:</p> <p>The response needs improvement. The student has not demonstrated an understanding of the complexities of the text. The response lacks specific examples from the essay to support the conclusion. The student needs to provide more details and explanations to clarify and extend their understanding beyond the literal.</p> <p>Mistakes/ Violations:</p> <ul style="list-style-type: none"> - The student needs to use proper capitalization and punctuation. - The response is too short and lacks details. The student needs to provide more information and examples from the essay to support their conclusion.

Fig. 10 Score variation example for Essay set-3

Essay_id = 8949, Human₁ Score=0, Human₂ Score=0, GPT-3.5-turbo Score=2
<u>Student Work</u>
The author concludes the story with that ending because he wanted to show the reader ‘if you just try, you might like it,’ and that’s why the girl said she would come back next spring when the snows melt so she can take the test again..
<u>GPT Response</u>
<p>Rubrics: Score 2: The response demonstrates a partial or literal understanding of the text. Addresses the demands of the question, although may not develop all parts equally Uses some expressed or implied information from the text to demonstrate understanding May not fully connect the support to a conclusion or assertion made about the text(s)</p> <p>Comments:</p> <ul style="list-style-type: none"> - The student has addressed the prompt and provided an answer to the question. - However, the response lacks depth and analysis of the text. - The student has not provided enough evidence from the text to support their answer. - The response could benefit from more elaboration and explanation. <p>Mistakes:</p> <ul style="list-style-type: none"> - The student has not made any major grammatical or spelling errors.

Fig. 11 Score variation example for Essay set-4

underestimated and the remaining 73 were overestimated by GPT. The possible reasons for the variations were same as of set-4 and set-5.

4.2 Gemini results for ASAP–AES

This section presents the results of employing the Gemini model for all the essay sets of ASAP–AES datasets. The results are summarized in subsequent sections presenting the detailed analysis for each essay set.

4.2.1 Essay set-1

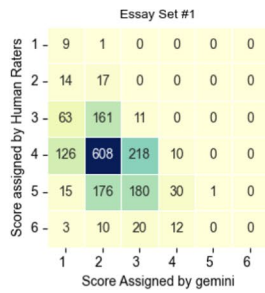
The essays for Essay set-1 were evaluated using the Gemini model and compared with human-assigned scores. Figure 12a illustrates the confusion matrix between the two sets of scores, while it provides an overview of the absolute differences in scores. Just as GPT struggled, the Gemini model could not perform promising results for this essay set and displayed significant variations in many cases. As shown in Fig. 12a, a majority of essays (608) were scored as 2 by Gemini where the human assessor assigned a score of 4. Further, notable discrepancies were observed for other score levels as well. For example, Gemini underestimated scores for 176 essays by assigning a 2 when human assessors were assigned a 5. Similarly, it assigned 3 to 180 essays where the human assessor assigned a score of 5. The absolute differences between the human and Gemini scores are summarized in Fig. 12b. The largest portion of essays (51.2%) exhibited a variation of 2 points, suggesting a significant deviation in these cases. Essays with a variation of 1 point accounted for 25.2%, while a variation of 3 points was observed for 19.1% of essays. Only 2.8% of essays had perfect agreement (no variation), and a negligible 0.2% showed variations of 4 or more points.

Overall, the Gemini model demonstrated a similar pattern to GPT. In most cases, Gemini underestimates the scores 2- to 3-point variations for majority of the essays. The possible reasons for such disagreements have already been discussed in Sect. 4.1.1.

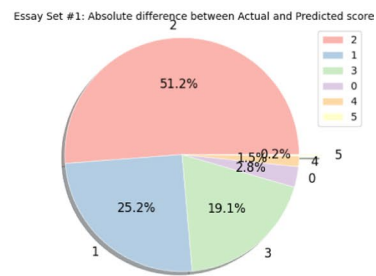
4.2.2 Essay set-2

The comparison between the scores of the human assessors and Gemini is shown in Fig. 12c for the writing implication domain. The results indicate that Gemini produced promising results and predicted scores were closer to human assessors, with a maximum variation of point 1 for majority of the essays. For instance, the Gemini and the human assessors both assigned a score of 4 to a large number of essays ($n = 497$), indicating a variation of 0. Similarly, 274 essays were assigned a score of 3 by human and Gemini. Furthermore, 327 essays were assigned a score of 4 by Gemini, where the human assessors' score was 3, indicating the variation of point 1. This implies that majority of the essays had the full agreement of the variation of point 1 for the writing implication domain.

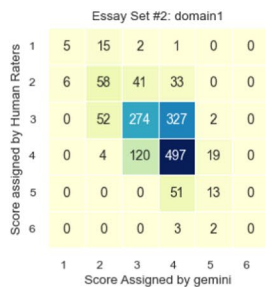
In addition, the results for language conventions are presented in Fig. 12e. A similar pattern can be observed for language convention, with a variation of point 1 or 0 for most of the essays. However, a few number of essays showed a variation of point 2. As can be observed from Fig. 12e, 498 essays were assigned the score of 3 by Gemini, where the human assessor's score was 4, indicating the variations of point 1. Similarly, Gemini assigned a score of 2 to 278 essays, whereas a human had assigned a score of 3. However,



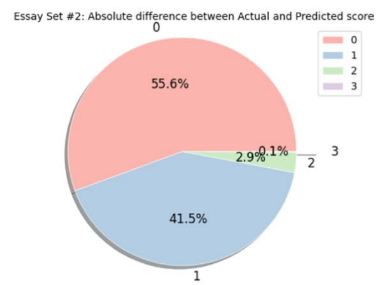
(a) CM - Set-1



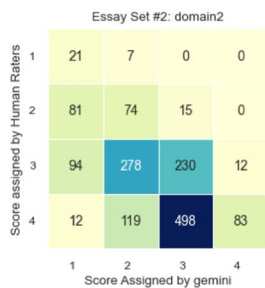
(b) Pie Chart - Set-1



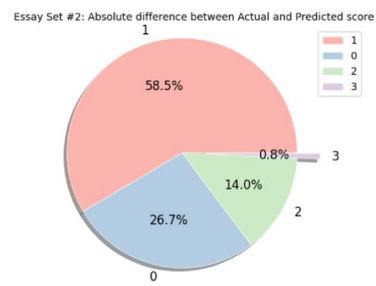
(c) CM - Set-2 Domain-1



(d) Pie Chart - Set-2 Domain-1



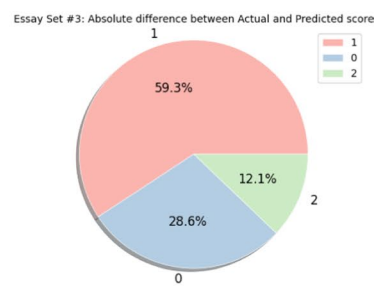
(e) CM - Set-2 Domain-2



(f) Pie Chart - Set-2 Domain-2



(g) CM - Set-3



(h) Pie Chart - Set-3

Fig. 12 Gemini results for ASAP-AES dataset (EssaySet 1, 2, and 3)

it is noticeable that a variation of 2 was also observed for a significant number of essays. Further, the variation of point 2 holds a great impact on the score range of 1–4.

The general variations between the scores of the human assessors and Gemini are illustrated in Fig. 12d for the writing implication tasks. Results indicate that Writing implications scores assigned by Gemini were closer to human assessors for majority of the essays, with a maximum variation of point 1. Only 2.9% and 0.1% essays exhibit a

variation of 2 and 3 points, respectively. In contrast, the scores for language convention showed a variation of 2 and 3 points for 14.0% and 0.8% respectively, as illustrated in Fig. 12f. Overall, it was observed that Gemini yielded better results for the writing implications domain as compared to the language convention domain. The possible reasons for better performance in Writing implications as compared to language conventions are already discussed in Sect. 4.1.2.

4.2.3 Essay set-3

This set consists of SDR essays that were written in response to the story “ROUGH ROAD AHEAD: Do Not Exceed Posted Speed Limit.” The students were tasked to explain how the characteristics of the environment affect the cyclist. The results presented in Fig. 12g indicate that the Gemini model achieved promising results for these SDR essays, with a maximum variation of 1 point for most of the essays. For example, Gemini and the human assessors both assigned a score of 2 to 408 essays, indicating a variation of 0. Furthermore, 173 essays were assigned a score of 1 by Gemini, whereas the human assessors’ score was 2, indicating a variation of 1 point.

In general, 28.6% and 59.3% of essays exhibit variations of 0 and 1, respectively, as illustrated in Fig. 12h. However, 12.1% of the essays exhibit variations of 2 points. Specifically, Gemini assigned a score of 0 to 16 essays, whereas human assessors assigned a score of 2. Similarly, Gemini assigned a score of 1 to 77 essays, whereas human assessors scored them as 3. These instances indicate that a total of 93 essays exhibited a variation above 1. To better understand these cases, a thorough analysis was conducted to investigate the potential factors contributing to the observed variations.

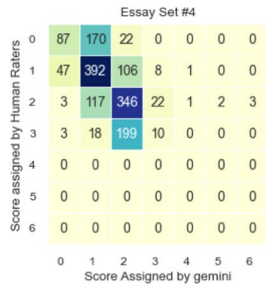
4.2.4 Essay set-4

The results for Essay set-4 are presented in Fig. 13a. It can be observed that the Gemini model achieved promising results for this set with accurately predicting the score for most of the essays. For example, Gemini and the human assessors both assigned a score of 1 to 392 essays, indicating a variation of 0. Similarly, 346 essays were assigned a score of 2 by both. Furthermore, 199 essays were assigned a score of 2 by Gemini, where the human assessors’ score was 3, indicating a variation of 1 point.

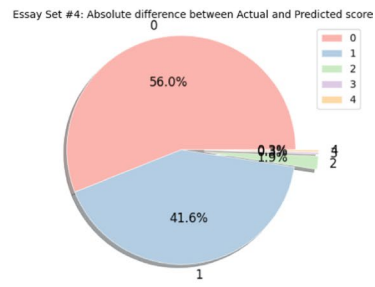
In general, a majority of essays showed an exact match with humans. 56.0% and 41.6% of essays exhibit variations of 0 and 1, respectively, as illustrated in Fig. 13b. However, only 2.4% of the essays exhibit variations above 1. This shows the prominent agreement between the human assessor and Gemini. It was observed that Gemini was very close to human assessors. However, it puts more emphasis on Grammar and other language convention. Hence, it provided comparative lower score as compared to human for some of the essays.

4.2.5 Essay set-5

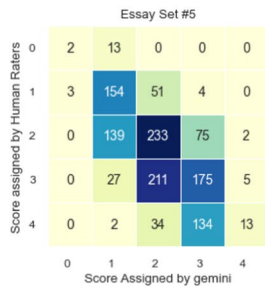
The results for Essay set-5 are shown in Fig. 13c. It can be observed that the Gemini model achieved strong performance for this set, closely aligning with human assessors for most essays. For instance, both Gemini and human raters assigned a score of 2 to 233 essays, indicating a perfect match (variation of 0). Similarly, 54 essays received a score of 1 from both raters.



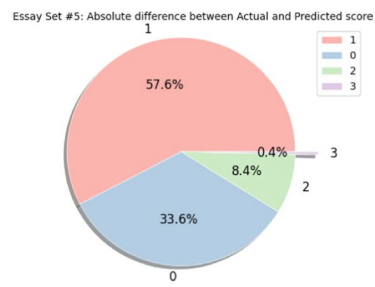
(a) CM - Set-4



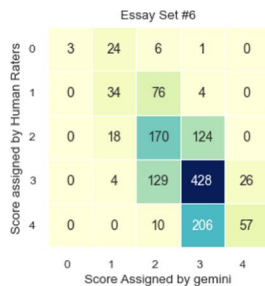
(b) Pie Chart - Set-4



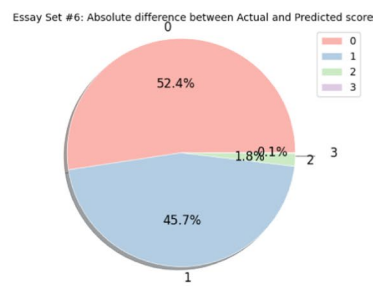
(c) CM - Set-5



(d) Pie Chart - Set-5



(e) CM - Set-6



(f) Pie Chart - Set-6

Fig. 13 Gemini results for ASAP–AES dataset (EssaySet 4, 5, and 6)

However, variations of 1 point were observed in certain cases. For example, Gemini assigned a score of 2 to 211 essays, whereas human raters assigned a score of 3, indicating a variation of 1. Likewise, 139 essays were scored as 1 by Gemini, while human raters were assigned a score of 2. In contrast, there were minimal instances of higher variations (greater than 1), highlighting the close agreement between Gemini and human raters.

Overall, a majority of the essays exhibited a variation of 1 with human scores. 57.6% of essays showed a variation of 1, while 33.6% exhibited a variation of 0. Notably, only 8.8% of essays demonstrated variations greater than 1 as reported in Fig. 13d. This demonstrates a high degree of alignment between Gemini and human assessments. However, similar to other sets, Gemini tended to focus more heavily on grammar and language conventions, leading to slightly lower scores for some essays compared to human ratings.

4.2.6 Essay set-6

The Essay set-6 results are presented in Figs. 13e, f. These figures indicate that the Gemini model achieved strong alignment with human raters for the majority of essays. For instance, both Gemini and human raters assigned a score of 3 to 428 essays,

demonstrating a perfect match (variation of 0). Similarly, 170 essays were scored as 2 by both raters.

However, certain variations were observed in certain cases. For example, Gemini assigned a score of 2 to 129 essays, whereas human raters assigned a score of 3, reflecting a variation of 1. Similarly, 124 essays received a score of 2 from human raters but were scored as 3 by Gemini, also showing a variation of 1. There were fewer instances of higher variations, such as cases where Gemini assigned a score of 4 to essays rated 3 or 2 by humans.

The pie chart (Fig. 13f) provides a breakdown of the absolute differences between Gemini and human scores. It shows that 52.4% of essays exhibited a variation of 0, indicating exact agreement. Furthermore, 45.7% of the essays had a variation of 1, while only 1.8% and 0.1% exhibited variations of 2 and 3, respectively. These results highlight the strong agreement between Gemini and human ratings, with a very small proportion of essays displaying significant deviations.

In summary, the majority of essays were rated identically by Gemini and human raters, demonstrating the model's capability to emulate human scoring. Nonetheless, similar to other essay sets, Gemini's scoring occasionally diverged slightly due to its emphasis on linguistic features such as grammar and syntax. This tendency may explain why some essays received slightly lower scores from Gemini compared to human ratings. Overall, the results underscore the model's robustness and consistency in essay scoring for this set.

4.3 Statistical analysis of results

QWK is considered the benchmark metric in AES research due to its ability to quantify inter-rater agreement while accounting for the ordinal nature of the scores and imposing higher penalty on larger deviations than smaller ones [2, 25, 59–62]. As a result, the QWK is particularly appropriate to use in the tasks where the scoring rubric describes a series of ordered categories rather than continuous interval data [63]. Recent LLM-based AES studies have consistently employed QWK as the primary evaluation metric [19, 20, 64], establishing it as the field standard for evaluating agreement between human raters and AES systems. QWK values range from -1 to 1 , where values closer to 1 indicate higher agreement, 0 represents chance agreement, and negative values indicate systematic disagreement. We adopted the established interpretation framework from educational measurement literature [65, 66], as presented in Table 4.

It should be noted that QWK functions as an effect-size measure of inter-rater agreement rather than a statistical hypothesis test. Consequently, it does not provide p-values or confidence intervals in the traditional hypothesis test framework [67]. The magnitude of the QWK coefficient is interpreted directly as the degree of agreement between raters.

Table 4 Interpretation of quadratic weighted kappa (QWK) scores

QWK range	Agreement level
< 0.20	Poor agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

To evaluate the reliability of LLM-based assessment, we computed QWK values across all rater combinations using scores assigned by Human Assessor 1 (HA_1), Human Assessor 2 (HA_2), Human Assessor resolved (HA_r), GPT, and Gemini. Hence, to contextualize our findings, we employed two key comparisons:

1. Human inter-rater reliability (HA_1 versus HA_2): This establishes the baseline agreement between expert human assessors
2. LLM versus human agreement (HA_r versus GPT/Gemini): This evaluates whether LLMs achieve agreement levels comparable to human inter-rater reliability

The computed QWK values are given in Table 5. When comparing LLM versus HA_r approach from HA_1 versus HA_2 , we can infer that the difference in the variability between language model and human scores is similar to the natural variability that is reported among trained human assessors. This comparison provides a concrete point of reference in the determination of the performance of language models in the scoring of essays.

4.3.1 QWK scores for ASAP–AES

- **HA_1 versus HA_2** The QWK scores between human assessors across all essay sets in the ASAP–AES dataset show consistently high agreement, ranging from 0.72 to 0.85. This demonstrates strong reliability among human assessors and indicates their high consistency in scoring essays within the dataset.
- **HA_r versus GPT** For comparing the scores of human assessors with GPT, we used the human assessor resolved score (HA_r) that was already reported in the ASA-AES dataset. The QWK scores for HA_r versus GPT indicate varying levels of agreement across essay sets:
 1. Low Agreement: Set1 shows inferior agreement (0.07), and Set2-d2 and Set3 exhibit slight-to-moderate agreement (0.25), highlighting GPT’s limitations in aligning with human scoring in these sets.
 2. Moderate Agreement: Moderate QWK scores for Set2-d1 (0.48), Set4 (0.45), and Set5 (0.46) indicate partial alignment, suggesting GPT performs better when scoring criteria align with its strengths.
 3. Best Performance: Set6 achieves the highest QWK (0.54), reflecting moderate-to-substantial agreement and better alignment with human assessors.

Table 5 QWK scores for human assessors, GPT, and Gemini

Dataset		HA_1 versus HA_2	HA_r versus GPT	HA_r versus Gemini
ASAP–AES	Set1	0.72	0.07	0.10
	Set2-d1	0.81	0.48	0.53
	Set2-d2	0.80	0.25	0.36
	Set3	0.77	0.25	0.38
	Set4	0.85	0.45	0.55
	Set5	0.75	0.46	0.58
	Set6	0.78	0.54	0.63
	Average	0.78	0.36	0.45
LA–AES	–	–	0.29	0.43

Note: HA_1 versus HA_2 represent two Human Assessors. HA_r refers to a resolved human assessor score

Overall, GPT achieves fair-to-moderate agreement, with inconsistent performance across sets, indicating a need for further refinement.

- **HA_r versus Gemini** For comparing the scores of human assessors with Gemini, we used the HA_r from the ASAP–AES dataset. The QWK scores for HA_r versus Gemini demonstrate higher agreement compared to GPT, showing Gemini’s improved alignment with human assessors across essay sets:
 1. **Moderate Agreement:** Gemini exhibits moderate agreement for Set1 (0.10), Set2-d2 (0.36), and Set3 (0.38). These scores, although better than GPT, highlight areas where Gemini could be further refined to match human judgment.
 2. **Substantial Agreement:** For Set2-d1 (0.53), Set4 (0.55), and Set5 (0.58), Gemini demonstrates substantial agreement with human raters. These results indicate that Gemini aligns well with human assessors when the scoring criteria are clearly defined and consistent with the model’s strengths.
 3. **Best Performance:** Set6 shows the highest agreement for Gemini with a QWK score of 0.63. This reflects substantial alignment with human raters and indicates that Gemini performs exceptionally well in scoring essays for this set.

Overall, Gemini consistently achieves better agreement with human assessors compared to GPT, particularly for essay sets where language conventions and structural elements are well-defined. However, there remains room for improvement in sets with more nuanced scoring criteria to achieve even closer alignment with human raters.

4.3.2 QWK scores for LA–AES

- **Human versus GPT** In the LA–AES dataset, GPT achieves a QWK score of 0.29. This score indicates limited agreement with human raters, similar to its performance in the ASAP–AES dataset. The lower alignment suggests that GPT struggles to replicate human grading patterns in this dataset as well.
- **Human versus Gemini** Gemini achieves a QWK score of 0.43 in the LA–AES dataset, outperforming GPT. While this score demonstrates a moderate level of agreement with human raters, it is still significantly lower than the inter-rater reliability observed among human assessors in the ASAP–AES dataset. Gemini’s performance reflects its relatively better ability to mimic human grading, though it also highlights the challenges of achieving human-like consistency in automated scoring.
- **Model Performance Comparison** Gemini-Pro’s superior performance over GPT–3.5-turbo could have happened due to two reasons. The first reason is that Gemini-Pro has more recent instruction-tuning, which allows it to adhere to more complicated rubric instruction and create more aligned assessments. The second is Gemini’s capability to capture longer contexts, as Table 3 indicates that its number of token-limit failures is lower. This context handling is better to allow Gemini to work with full-sized rubrics and essay text in a more effective manner, resulting in increased concurrence by the human assessor.

4.3.3 Key findings for ASAP–AES

We determine the extent of alignment between the scores assigned by human assessors and LLMs (GPT and Gemini) using QWK. When comparing the scores of LLMs with HA_r , the results are as follows:

- **GPT Performance:** The QWK scores for GPT show significant variation across essay sets (as discussed in Sect. 4.3.1). While these results validate GPT’s potential as an essay scorer, the inconsistency across datasets suggests that the scoring of GPT models is also dependent on prompt and rubric properties, thus requiring additional adjustments to achieve greater consistency between the scoring of humans and GPT models.
- **Gemini Performance:** The QWK scores for Gemini indicate stronger alignment with human scores compared to GPT (as discussed in Sect. 4.3.1). These results indicate that there is no significant difference between Gemini and human scores in most cases. Gemini consistently outperforms GPT, particularly in sets with clear evaluation rubrics.

Hence, we validate through statistical testing that Gemini demonstrates greater potential for consistent and reliable essay scoring compared to GPT.

Additionally, to determine the level of alignments between human assessors, the QWK scores in Table 5, HA_1 versus HA_2 were observed, and they exhibit consistently high agreement (QWK scores ranging from **0.72 to 0.85**). It supports the conclusion that the level of agreement between human assessors is “Substantial”. This highlights the strong reliability and consistency of human scoring.

One possible explanation for the lower QWK scores for Gemini versus HA_r as compared to HA_1 versus HA_2 is the presence of human biases and errors in manual scoring. It was observed that Gemini assigned a bit lower score to essays as compared to humans. In most of the cases, the variation was 1 point. This indicates that Human assessors are subject to cognitive biases, such as halo effects or leniency biases, which can lead to variability in scoring. Additionally, the sequential order of essay scoring may influence scoring rigor, with earlier essays being more closely aligned with rubrics than later ones. Additionally, Gemini’s stronger alignment with human scores is possibly due to its enhanced evaluation framework and contextual understanding. This objectivity highlights areas that human assessors might miss, offering students actionable insights for refinement. In conclusion, Gemini demonstrates greater consistency and alignment with human raters, making it a more effective tool for essay scoring. Both systems highlight the potential for reducing human biases and improving the objectivity of essay scoring through automation.

4.3.4 Key findings for LA–AES

The results for the LA–AES dataset, as shown in Table 5, highlight the following key findings to compare the results of humans with LLMs i.e., HA_r versus GPT and HA_r versus Gemini:

- GPT exhibits weaker alignment with human assessors, with its QWK score remains 0.29, indicating limited reliability as a standalone scoring system for this dataset.
- The QWK score for Gemini when compared with HA_r is 0.43, reflecting moderate agreement. This performance is notably stronger than GPT (that is 0.29), showcasing

Gemini's improved ability to align with human scoring criteria. Despite moderate agreement, the score indicates that there are still measurable differences, suggesting further refinement is needed. For this dataset, it is concluded that there is no significant difference in scoring for Gemini in many cases.

These findings validate that Gemini is a more effective essay scorer for the LA–AES dataset, with better alignment to human scoring than GPT. However, both systems highlight the potential for improving objectivity and consistency in automated essay scoring.

4.4 Evaluation on real-life dataset

The proposed approach to assigning the score and providing detailed feedback to the students was evaluated in a real-life scenario to investigate human assessor subjectivity and fatigue rather than serving as a large-scale performance benchmark. Specifically, we conducted the evaluation with the O-Levels class of the Sukkur IBA community college. We collected 21 responses written by students from Google Classroom. The subject teacher provided valuable cooperation and assistance throughout the process. The evaluation encompassed a diverse range of essays, covering various topics and writing styles. Students and two domain experts, including the subject teacher (Human Assessor1, HA₁) and an educator from another class (Human Assessor2, HA₂), participated in the evaluation process to provide a comprehensive perspective on the quality of the proposed approach.

In the evaluation process, the collected 21 essays were initially assessed by the subject teacher, and anonymity was not maintained during this stage. Next, we anonymized the student responses by assigning a unique ID to the essay to introduce anonymity. The IDs of the essays followed the same sequence in which the subject teacher assessed the responses to track assessor fatigue over time. For example, the essay assessed at the beginning was assigned ID e001, and the essay assessed at the end was assigned ID e021. The comparison is depicted in Fig. 14 for the scores assigned by LLMs and human assessors, respectively. It can be observed that most of the essays that were assessed earlier show a minimal variation. However, the responses assessed later show a high rate of variation, demonstrating human agreement with LLMs (and with other human assessors) significantly declined as the assessor became fatigued, whereas the LLM remained

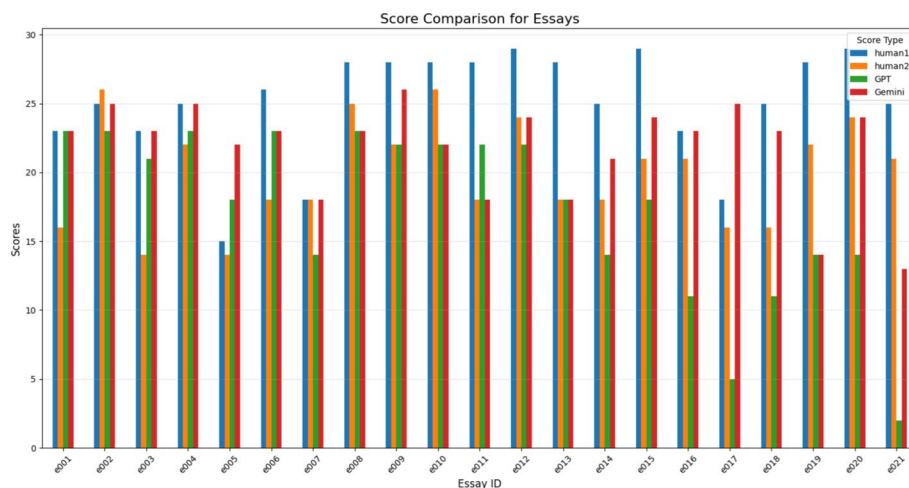


Fig. 14 Comparison of LLMs with human assessors for O-Levels

objective. In the second phase of assessment, anonymized responses were assessed by another assessor who has expertise in the field. The results for this phase were astonishing; a strong disagreement was found between the HA₁ and HA₂, as shown in Fig. 14.

This evaluation provided insights into the implications of the LLMs in assigning scores and providing detailed feedback to students to improve their writing skills. The results produced from this evaluation are summarized in Table 6. It was evident from the evaluation that Gemini exhibited a pattern similar to the ASAP–AES and LA–AES standard datasets. Furthermore, the key findings and comparisons between human assessments and those of LLMs, derived from the evaluation results, are as follows:

- **HA₁ and HA₂ assessment:** The results presented in Fig. 14 reveal a significant variation between the scores assigned by HA₁ and HA₂. For example, HA₁ assigned scores of 23, 26, 28, and 25 to essays e003, e006, e011, and e018, respectively, while HA₂ assigned scores of 14, 18, 18, and 16 to the same essays. This demonstrates a notable level of disagreement between the two human assessors. It indicates that human assessment is subjective; they might have a large number of essays to evaluate within a limited timeframe, which can lead to fatigue or rushing through the assessment process. Hence, some important details may be overlooked or ignored by human assessors. Further, it is almost impossible to provide detailed feedback to every student, highlighting their mistakes and suggesting corrections.
- **HAs and LLMs:** Both LLMs demonstrate a good level of agreement with HA₂ but exhibit significant variation when compared to HA₁. The QWK scores indicate minimal agreement between the LLMs and HA₁, with very low values: 0.13 for GPT and 0.01 for Gemini. In contrast, better QWK values were observed for HA₂, with 0.3 for GPT and 0.19 for Gemini (as reported in Table 7). Notably, GPT outperformed Gemini in terms of QWK scores, likely due to the use of the Weighted Kappa metric, which emphasizes the magnitude of differences. Given that the scoring scale for

Table 6 Scores assigned by LLM and human assessors for O-levels class

Essay ID	Essay type	HA ₁	HA ₂	GPT	Gemini
e001	Report Writing	23	16	23	23
e002	Informal letter	25	26	23	25
e003	Report Writing	23	14	21	23
e004	Report Writing	25	22	23	25
e005	Descriptive	15	14	18	22
e006	Report Writing	26	18	23	23
e007	Descriptive	18	18	14	18
e008	Report Writing	28	25	23	23
e009	Descriptive	28	22	22	26
e010	Descriptive	28	26	22	22
e011	Descriptive	28	18	22	18
e012	Narrative	29	24	22	24
e013	Descriptive	28	18	18	18
e014	Narrative	25	18	14	21
e015	Narrative	29	21	18	24
e016	Informal letter	23	21	11	23
e017	Informal letter	18	16	5	25
e018	Informal letter	25	16	11	23
e019	Descriptive	28	22	14	14
e020	Narrative	29	24	14	24
e021	Informal letter	25	21	2	13

Table 7 QWK scores for pairwise comparisons for O-levels class

Comparison	QWK
HA ₁ versus HA ₂	0.30
HA ₁ versus GPT	0.13
HA ₁ versus Gemini	0.01
HA ₂ versus GPT	0.30
HA ₂ versus Gemini	0.19

this assessment ranges from 2 to 30, these differences are accentuated. However, the graph also shows that Gemini produced results closely aligned with HA₂. This underscores the fact that LLM assessments tend to be more objective compared to human assessors. In conclusion, high variances between HA₁ and GPT were observed, with GPT's scores being closer to HA₂, shows that human assessors may be susceptible to personal preferences or leniency biases. Human assessors might have considered various factors unintentionally while assessing the work, such as class performance, previous grades of students, and class participation. In the results, some of the students may get the advantage of leniency bias because of their good grades and participation in the class.

- Most of the essays that were assessed earlier by HA₁ show a minimal variation when comparing the results with LLMs. However, the responses assessed later show a high rate of variation. This is possibly due to the halo effect and fatigue. In this scenario, Human assessors might have made a judgment by just reading the first few sentences and skimming the rest of the part.

Hence, it is concluded that it is beneficial to consider using automated systems like GPT-3.5-turbo and Gemini in conjunction with human assessments to mitigate the potential biases introduced by human assessors. This combination can help minimize inconsistencies and provide a more standardized evaluation process. Additionally, establishing clear evaluation criteria and guidelines can help minimize the impact of personal preferences and biases when assessing student work.

5 Conclusion and future research direction

In this study, we proposed an approach for AES along with feedback provision by utilizing the GPT-3.5-turbo and Gemini-pro models. We evaluated our approach on the ASAP-AES and LA-AES benchmark datasets along with the real-life scenarios of the O-Levels class. Our proposed approach yields promising results in terms of predicting scores and providing comprehensive feedback for essays following the given rubric guidelines. It can deal with various domains and different types of rubrics. However, the quality of rubrics can affect the performance of the system to a high extent. For example, the word 'detailed' is very subjective and can be considered differently by different people. The performance model was a bit lower for the instances where such terms were mentioned in the rubric guidelines. So, instead of mentioning the word 'detailed' in rubrics, the number of words should be specified, such as "between 200 and 300 words". Hence, the rubric must be clearly formatted and follow the standard language for each aspect of writing. In addition, the analysis of the O-Levels category showed that the human assessors are susceptible to diverse cognitive and assessor-specific biases such as severity or leniency bias, personal preferences, fatigue, and the halo effect. The outcomes indicate a massive variance of the two human assessors of identical rubric guidelines.

According to our results, the scoring using LLM is more objective in reducing such cognitive biases than the scoring of human assessors.

Although this study has focused on using GPT-3.5-turbo and Gemini-pro, it is worth mentioning that these models have been superseded by more advanced versions, such as GPT-4, Gemini 1.5, Claude 3, and Llama-3. Therefore, future research could explore whether these newer models improve performance and efficiency in evaluating student essays. In addition, future research could also evaluate the proposed approach across various other subjects beyond essay writing, such as programming assessment, scientific experiments, and mathematical equations. Finally, we emphasize that our definition of bias refers specifically to human cognitive and assessor biases, not demographic or social biases. A separate fairness analysis would therefore be valuable to investigate potential demographic biases related to variables such as gender, ethnicity, and socio-economic background, ensuring fair evaluation across diverse student populations.

Acknowledgements

We sincerely thank Mrs. Sanam Zaryab, English Lecturer (O-Levels) at Sukkur IBA Community College, and Mr. Usama Abdul Rehman, English Lecturer at Sukkur IBA University, for generously providing classroom data and sharing their academic expertise. Their contributions were instrumental to this study on automated essay scoring and human assessment alignment.

Author contributions

Nimra Mughal: conceived the idea, performed all experiments related to the study, and reported results. Ali Shariq Imran: contributed to designing the research methodology and supervision of the project. Sher Muhammad Daudpota: contributed to designing the research methodology and supervised the experiments. Zenuk Kastrati: contributed to designing the research methodology and supervised the project. Waheed Noor: contributed to designing of the research methodology and provided a review of the paper. All authors read and approved the final manuscript.

Funding

Open access funding provided by Linnaeus University.

Data availability

The benchmark datasets (ASAP-AES and LA-AES) are publicly available. The real-world O-Level dataset and experimental code are available in our GitHub repository: <https://github.com/nimra16/Automated-essay-scoring-using-LLM.git>. A live demonstration of the system's graphical user interface is accessible at: https://nimra16.pythonanywhere.com/assessment_checker.

Declarations

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest.

Received: 18 September 2025 / Accepted: 6 February 2026

Published online: 18 February 2026

References

1. Nisbett RE, Wilson TD. The halo effect: evidence for unconscious alteration of judgments. *J Pers Soc Psychol.* 1977;35(4):250.
2. Ramesh D, Sanampudi SK. An automated essay scoring systems: a systematic literature review. *Artif Intell Rev.* 2022;55(3):2495–527.
3. Crossley SA. Linguistic features in writing quality and development: An overview. *J Writ Res.* 2020;11(3):415–43.
4. Kyle K, Crossley S. Assessing syntactic sophistication in L2 writing: a usage-based approach. *Lang Test.* 2017;34(4):513–35.
5. Maulud DH, Zeebaree SR, Jacksi K, Sadeeq MAM, Sharif KH. State of art for semantic analysis of natural language processing. *Qubahan Acad J.* 2021;1(2):21–8.
6. Hearst MA. The debate on automated essay grading. *IEEE Intell Syst Appl.* 2000;15(5):22–37.
7. Collins-Thompson K. Computational assessment of text readability: a survey of current and future research. *ITL-Int J Appl Linguist.* 2014;165(2):97–135.
8. Lim CT, Bong CH, Wong WS, Lee NK. A comprehensive review of automated essay scoring (AES) research and development. *Pertanika J Sci Technol.* 2021;29(3):1875–99.

9. Medsker LR, Jain L. Recurrent neural networks. *Des Appl*. 2001;5:64–7.
10. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput*. 2019;31(7):1235–70.
11. Ludwig S, Mayer C, Hansen C, Eilers K, Brandt S. Automated essay scoring using transformer models. *Psych*. 2021;3(4):897–915.
12. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. *AI Open*. 2022;3:111–32.
13. Xue J, Tang X, Zheng L. A hierarchical bert-based transfer learning approach for multi-dimensional essay scoring. *IEEE Access*. 2021;9:125403–15.
14. Koroteev M. Bert: a review of applications in natural language processing and understanding. arXiv preprint [arXiv:2103.11943](https://arxiv.org/abs/2103.11943). 2021.
15. Mayfield E, Black AW. Should you fine-tune bert for automated essay scoring? In: Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications. 2020. pp. 151–162.
16. Prabhu S, Akhila KSS. A hybrid approach towards automated essay evaluation based on bert and feature engineering. In: 2022 IEEE 7th international conference for convergence in technology (I2CT). 2022. pp. 1–4. <https://doi.org/10.1109/I2CT54291.2022.9824999>.
17. Zheng C, Huang L, Lin H, Guo Y, Huang L. Bert-based automatic scoring model for speech-oriented text modality. In: 2022 IEEE 2nd international conference on electronic technology, communication and information (ICETCI). IEEE; 2022. pp. 100–105.
18. Song Y, Zhu Q, Wang H, Zheng Q. Automated essay scoring and revising based on open-source large language models. *IEEE Trans Learn Technol*. 2024a;17:1880–90. <https://doi.org/10.1109/TLT.2024.3396873>.
19. Li W, Liu H. Applying large language models for automated essay scoring for non-native Japanese. *Human Social Sci Commun*. 2024;11(1):1–15.
20. Lee S, Cai Y, Meng D, Wang Z, Wu Y. Unleashing large language models' proficiency in zero-shot essay scoring. In: Findings of the association for computational linguistics: EMNLP 2024. 2024. pp. 181–198.
21. Blanchard D, Tetreault J, Higgins D, Cahill A, Chodorow M. Toefl11: A corpus of non-native English. *ETS Res Rep Ser*. 2013;2013(2):15.
22. Abraham A. Rule-based expert systems. *Handbook of measuring system design*. 2005.
23. Cowell RG, Dawid P, Lauritzen SL, Spiegelhalter DJ. Probabilistic networks and expert systems: exact computational methods for Bayesian networks. New York: Springer; 2007.
24. Ramalingam V, Pandian A, Chetry P, Nigam H. Automated essay grading using machine learning algorithm. In: Publishing IOP, editor. *Journal of physics: conference series*, vol. 1000. Bristol; 2018. p. 012030.
25. Ke Z, Ng V. Automated essay scoring: a survey of the state of the art. In: *IJCAI*, vol 19. 2019. pp. 6300–6308.
26. Baidoo-Anu D, Owusu Ansah L. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. Available at SSRN 4337484. 2023.
27. Nunes A, Cordeiro C, Limpo T, Castro SL. Effectiveness of automated writing evaluation systems in school settings: a systematic review of studies from 2000 to 2020. *J Comput Assist Learn*. 2022;38(2):599–620.
28. Connor U. Linguistic/rhetorical measures for international persuasive student writing. *Res Teach English*. 1990;67–87.
29. Parra GL, Calero SX. Automated writing evaluation tools in the improvement of the writing skill. *Int J Instr*. 2019;12(2):209–26.
30. Hussein MA, Hassan H, Nassef M. Automated language essay scoring systems: a literature review. *PeerJ Comput Sci*. 2019;5:208.
31. Ifenthaler D. Automated essay scoring systems. In: *Handbook of open distance and digital education*. New York: Springer; 2022. p. 1–15.
32. Lu C, Cutumisu M. Integrating deep learning into an automated feedback generation system for automated essay scoring. *Int Educ Data Min Soc*. 2021.
33. Attali Y, Burstein J. Automated essay scoring with e-rater® v.2. *J Technol Learn Assess* 2006;4(3).
34. Alpaydin E. Introduction to machine learning. Cambridge: MIT press; 2020.
35. Landauer TK. Automatic essay assessment. *Assess Educ Principles Policy Practice*. 2003;10(3):295–308.
36. Mahesh B. Machine learning algorithms-a review. *Int J Sci Res (IJSR) [Internet]*. 2020;9:381–6.
37. Cohen Y, Ben-Simon A, Hovav M. The effect of specific language features on the complexity of systems for automated essay scoring. 2003.
38. Mahana M, Johns M, Apte A. Automated essay grading using machine learning. *Mach. Learn Session* 2012;5.
39. Zhang Y, Jin R, Zhou Z-H. Understanding bag-of-words model: a statistical framework. *Int J Mach Learn Cybern*. 2010;1:43–52.
40. Heeman PA. Pos tags and decision trees for language modeling. In: 1999 joint SIGDAT conference on empirical methods in natural language processing and very large corpora. 1999.
41. Venezky RL. The structure of English orthography. In: *The structure of English orthography*. Berlin: De Gruyter Mouton; 2011.
42. Dargan S, Kumar M, Ayyagari MR, Kumar G. A survey of deep learning and its applications: a new paradigm to machine learning. *Arch Comput Methods Eng*. 2020;27:1071–92.
43. Chavez MR, Butler TS, Rekawek P, Heo H, Kinzler WL. Chat generative pre-trained transformer: why we should embrace this technology. *Am J Obstet Gynecol*. 2023;228(6):706–11.
44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A.N, Kaiser Ł, Polosukhin I. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30.
45. Tenney I, Das D, Pavlick E. Bert rediscovers the classical nlp pipeline. arXiv preprint [arXiv:1905.05950](https://arxiv.org/abs/1905.05950). 2019.
46. Dale R. Gpt-3: What's it good for? *Nat Lang Eng*. 2021;27(1):113–8.
47. Annepaka Y, Pakray P. Large language models: A survey of their development, capabilities, and applications. *Knowl Inf Syst* 2024;1–56.
48. Cohen V, Gokaslan A. Opengpt-2: Open language models and implications of generated text. *XRDS Crossroads ACM Mag Stud*. 2020;27(1):26–30.
49. Imran M, Almusharraf N. Google gemini as a next generation ai educational tool: a review of emerging educational technology. *Smart Learn Environ*. 2024;11(1):22.

50. Masalkhi M, Ong J, Waisberg E, Zaman N, Sarker P, Lee AG, Tavakkoli A. A side-by-side evaluation of llama 2 by meta with chatgpt and its application in ophthalmology. *Eye* 2024;1–4.
51. Meyer L, Dannecker A. Comparative analysis of generative ai models in educational exercise performance. In: EDU-LEARN24 Proceedings. IATED; 2024. pp. 5181–5190.
52. Floridi L, Chiriatti M. Gpt-3: its nature, scope, limits, and consequences. *Mind Mach.* 2020;30:681–94.
53. Lund BD, Wang T. Chatting about chatgpt: How may ai and gpt impact academia and libraries? *Library Hi Tech News.* 2023;40(3):26–9.
54. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. Chatgpt for good? On opportunities and challenges of large language models for education. *Learn Individ Differ.* 2023;103:102274.
55. Mizumoto A, Eguchi M. Exploring the potential of using an ai language model for automated essay scoring. *Res Methods Appl Linguist.* 2023;2(2):100050.
56. Latif E, Zhai X. Large language models and automated essay scoring of English language learner writing: insights into validity and reliability. *Comput Educ Artif Intell.* 2024;6:100213. <https://doi.org/10.1016/j.caeai.2024.100213>.
57. Amin T, Aadil F, Awan KM, Lim S, et al. Enhancing essay scoring: an analytical and holistic approach with few-shot transformer-based models. *IEEE Access.* 2025.
58. Connor U, Lauer J. Understanding persuasive essay writing: linguistic/rhetorical approach. *Text-Interdiscip J Study Discourse.* 1985;5(4):309–26.
59. Williamson DM, Xi X, Breyer FJ. A framework for evaluation and use of automated scoring. *Educ Meas Issues Pract.* 2012;31(1):2–13.
60. Shermis MD, Hamner B. Contrasting state-of-the-art automated scoring of essays: Analysis. In: Annual National Council on Measurement in Education Meeting. 2012.
61. Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading ESOL texts. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011. pp. 180–189.
62. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull.* 1968;70(4):213–20.
63. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Measur.* 1973;33(3):613–9.
64. Song Y, Zhu Q, Wang H, Zheng Q. Automated essay scoring and revising based on open-source large language models. *IEEE Trans Learn Technol.* 2024b;17:1880–90.
65. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–74.
66. Altman DG. *Practical statistics for medical research.* London: Chapman and Hall/CRC; 1990.
67. Gwet KL. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters.* 4th ed. Piedmont: Advanced Analyticsn, LLC; 2014.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.