Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1335

ChatGPT vs. Machine Learning: Assessing the Efficacy and Accuracy of Large Language Models for Automated Essay Scoring

Youngwon Kim
Harvard University

Reagan Mozer
Bentley University

Shireen Al-Adeimi
Michigan State University

Luke Miratrix
Harvard University

Automated Essay Scoring (AES) is a critical tool in education that aims to enhance the efficiency and objectivity of educational assessments. Recent advancements in Large Language Models (LLMs), such as ChatGPT, have sparked interest in their potential for AES. However, comprehensive comparisons of LLM-based methods with traditional machine learning (ML) methods across different assessment contexts remain limited. This study compares the efficacy of LLMs with supervised ML algorithms in assessing both categorical essay opinions and continuous writing quality scores. Using two distinct datasets–argumentative essays from 4th-7th graders about iPad usage in schools, and persuasive essays from 10th graders on censorship in libraries–we systematically assess the performance of ChatGPT compared to four tree-based ML algorithms trained on extensive statistical text features. Our findings show that while LLMs perform well in essay classification tasks, ML methods consistently outperform LLMs in predicting writing quality. We highlight the importance of prompting and fine tuning techniques in LLM-based scoring, along with the strengths and limitations of both approaches. We also discuss the potential of LLMs to enhance AES in educational settings while underscoring the continued importance of human oversight in evaluating complex writing skills. Overall, this study demonstrates the complementary strengths of different approaches to AES, providing guidance for researchers and educators interested in leveraging LLMs in educational assessment.

VERSION: November 2025

Suggested citation: Kim, Youngwon, Reagan Mozer, Shireen Al-Adeimi, and Luke Miratrix. (2025). ChatGPT vs. Machine Learning: Assessing the Efficacy and Accuracy of Large Language Models for Automated Essay Scoring. (EdWorkingPaper: 25-1335). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7vj9-5y53

ChatGPT vs. Machine Learning: Assessing the Efficacy and Accuracy of Large Language Models for Automated Essay Scoring

Youngwon Kim[1], Reagan Mozer[2*], Shireen Al-Adeimi[3], Luke Miratrix[1]

[1*]Graduate School of Education, Harvard University, Street, Cambridge, 02138, MA, USA.
[2]Department of Mathematical Sciences, Bentley University, 175 Forest Street, Waltham, 02452, MA, USA.
[3]College of Education, Michigan State University, Street, City, 610101, State, USA.

*Corresponding author(s). E-mail(s): rmozer@bentley.edu;
Contributing authors: youngwonkim.ywk@gmail.com;
aladeimi@msu.edu; lmiratrix@g.harvard.edu;

Abstract

Automated Essay Scoring (AES) is a critical tool in education that aims to enhance the efficiency and objectivity of educational assessments. Recent advancements in Large Language Models (LLMs), such as ChatGPT, have sparked interest in their potential for AES. However, comprehensive comparisons of LLM-based methods with traditional machine learning (ML) methods across different assessment contexts remain limited. This study compares the efficacy of LLMs with supervised ML algorithms in assessing both categorical essay opinions and continuous writing quality scores. Using two distinct datasets–argumentative essays from 4th-7th graders about iPad usage in schools, and persuasive essays from 10th graders on censorship in libraries–we systematically assess the performance of ChatGPT compared to four tree-based ML algorithms trained on extensive statistical text features. Our findings show that while LLMs perform well in essay classification tasks, ML methods consistently outperform LLMs in predicting writing quality. We highlight the importance of prompting and fine-tuning techniques in LLM-based scoring, along with the strengths and limitations of both approaches. We also discuss the potential of LLMs to enhance AES in educational settings while underscoring the continued importance of human

oversight in evaluating complex writing skills. Overall, this study demonstrates the complementary strengths of different approaches to AES, providing guidance for researchers and educators interested in leveraging LLMs in educational assessment.

Keywords: Automated scoring, Educational assessment, Machine learning, ChatGPT

1 Introduction

Evaluating student essays is a vital component of education, as it enables educators and researchers to gauge students' critical thinking skills, writing proficiency, and comprehension. The evaluation process has long relied on human assessors, due to the complex nature of writing, a productive task that demands an in-depth assessment of various elements (Hussein et al. 2019). These elements include content, writing quality, adherence to guidelines, coherence, and grammatical precision, all while considering the individual circumstances and academic skills of students. Further, manual evaluation requires access to well-trained and experienced human graders who are able to sustain a high level of mental engagement throughout the scoring process. Thus, human scoring efforts in educational and research settings can be labor-intensive, time-consuming, and costly (Ramesh and Sanampudi 2022).

These inherent challenges have spurred exploration into alternative approaches, with automated essay scoring (AES) emerging as a promising solution (Shermis and Hamner 2013). AES refers to a computer-based assessment approach to evaluate and score written work (Ramesh and Sanampudi 2022; Dikli 2006; Kumar and Boulanger 2020). It aims to overcome the challenges of manual essay grading, including time constraints, cost, potential inconsistencies among human graders, and scalability issues (Sevcikova 2018; Ramineni and Williamson 2013). AES has found widespread use in supporting the measurement and analysis of text data across various educational settings. However, the promise of entirely eliminating human coding efforts remains impractical in most real-world educational scenarios, largely due to the high costs and the absence of universally applicable tools (Weegar and Idestam-Almquist 2024).

The release of ChatGPT at the end of 2022 has catalyzed a wave of research and experimentation in the field of AES, and is beginning to reshape the practical landscape of its use (Altamimi 2023; Xia et al. 2024). As one of the leading large language models (LLMs), ChatGPT exhibits remarkable language understanding, mimicking the judgment of educators in assessing the quality of student writing, and potentially replacing human labor and judgment. Educators and researchers are increasingly interested in exploring the applications and methodologies of LLMs for essay scoring, with the aim of utilizing their capabilities to enhance the efficiency and objectivity of the grading process. Despite growing interest, the full potential of LLMs like ChatGPT for AES remains understudied compared to machine learning (ML) methods, which have long been a cornerstone of AES and typically rely on statistical feature extraction.

This research seeks to address this gap by posing a critical question: Do LLMs outperform established ML methods for AES, and can even better results be achieved by

combining their strengths? In the sections that follow, we first present a brief overview of AES, examining the use of both ML and state-of-the-art LLMs in this context. Through a series of case studies, we then compare the performance of LLM-based scoring with traditional supervised ML methods and explore their effectiveness for various types of outcomes, including both continuous scores and categorical classifications.

The present study aims to make several contributions to the field. Our first contribution lies in providing insights into the utilization of LLMs and ML algorithms for essay grading in fully automated contexts. Our second contribution involves investigating how these approaches perform at predicting both continuous and categorical outcomes. Overall, this work adds to the ongoing discourse surrounding the integration of artificial intelligence in education and shed light on both the opportunities and challenges associated with automating complex cognitive tasks such as essay grading.

2 Background

Pioneering work in AES began in the 1960s with Ellis Page's Project Essay Grader (PEG), which focused on surface-level writing features (e.g., word count, average sentence length, number of connection words) and applied them in multiple regression analysis (Page 1994, 2003). Advancements in technology and statistical methods have since led to increasingly sophisticated AES models and active research on AES. Beyond surface-level features, systems like the Intelligent Essay Assessor (IEA) incorporated content analysis with latent semantic analysis (Foltz et al. 1999). Tools like E-rater (1998), Intellimetric (2006), and Bayesian Essay Test Scoring System (BESTY, 2002) utilize natural language processing (NLP) to consider both style and content (Ramineni and Williamson 2013; Lim et al. 2021). These AES systems often rely on task-specific linguistic features and regression-based ML to evaluate writing quality (Shermis and Hamner 2013).

However, some argue that ML approaches cannot fully grasp the depth and complexity of essay assessment in that they are constrained to surface-level feature extraction and may fail to capture the multifaceted and profound attributes considered by human assessors during the actual grading process (Ramesh and Sanampudi 2022). To address this, attention-based deep learning techniques have been explored for AES (Dong et al. 2017; Taghipour and Ng 2016). The development of pre-training and fine-tuning methods has further advanced deep neural network-based NLP, improving contextual understanding of written essays (Peters et al. 2018). These advancements have substantially contributed to the development of LLMs like ChatGPT in 2022. Although LLMs demonstrate remarkable language generation capabilities, their direct application in AES remains an area for further investigation.

2.1 Machine Learning Models for Automated Essay Scoring

ML-based AES systems employ a structured pipeline to effectively evaluate and score written essays, thus relying on text features extracted from the essays (Xia et al. 2024). This process begins with generating a comprehensive grading rubric that defines the criteria and scoring levels for various aspects of writing (e.g., grammar, organization, and content relevance). Once the rubric is established, a representative set of essays

is hand-scored by human experts according to these guidelines. These coded essays serve as annotated data in the training set for the ML model. It is crucial to ensure this sample captures the range of topics, writing styles, and potential issues that texts might present (Lim et al. 2021).

Before extracting text features, pre-processing the text data is a critical step. This may involve tasks like tokenization, normalization, and removing stop words. Following pre-processing, the next step is to select and extract rubric-aligned features that effectively represent the essays' content and style. These features can include basic textual statistics like word count, lexical diversity, and grammatical accuracy, alongside more complex linguistic patterns that might indicate higher levels of reasoning or rhetorical skills (Mozer et al. 2024). A diverse set of numerical features is necessary to create a high-performing prediction model (Ramesh and Sanampudi 2022; Mizumoto and Eguchi 2023; Hussein et al. 2019). After training, the model's performance is evaluated using a separate validation set of essays not included during training. Performance metrics such as accuracy and kappa values determine how well the model predicts essay scores. Based on these results, further iterations might be necessary, potentially involving tuning hyperparameters, modifying features, or even redefining parts of the rubric.

In this study, we explore whether LLMs, with their deeper contextual understanding, can outperform or complement this feature-based ML approach, thereby providing educators with powerful alternatives for essay scoring.

2.2 Large Language Models for Automated Essay Scoring

LLMs represent a paradigm shift in AES research, as they utilize deep learning architectures such as transformers to process and generate natural language with human-like fluency and coherence (Nazir and Wang 2023). Unlike traditional ML-based AES systems, which extract linguistic and statistical features from text that are then used as inputs in a model to predict human-assigned scores, LLMs operate through text generation and contextual understanding. Their ability to evaluate writing holistically opens new avenues for automated assessment and feedback generation that more closely mirror the capabilities of a human grader (Xia et al. 2024). On the other hand, the use of LLMs for AES introduces new challenges including potential hallucinations and the perpetuation of inherent biases from training data (Božić and Poola 2023; Nazir and Wang 2023; Kocoń et al. 2023).

2.2.1 Scoring Process

The AES workflow using LLMs differs considerably from traditional ML-based approaches. While both approaches begin with clearly defining the specific aspect of writing to be assessed (e.g., argumentation quality or organization), the subsequent steps diverge. In LLM-based scoring, a structured prompt based on established scoring rubrics serves as the primary input to the model and guides its response generation. This eliminates the need for extensive text pre-processing and feature engineering required by ML models but introduces new challenges related to prompt design and variability in model outputs (Ekin 2023; Ozdemir 2023; Heston and Khun 2023).

For example, LLMs can produce varying responses to the same input depending on hyperparameters such as response temperature (which controls response randomness). To mitigate output variability, one strategy is to run the same prompt multiple times for each essay and aggregate the results. Integrating few-shot prompt techniques that include specific examples within the prompt itself can also be beneficial (Heston and Khun 2023). Furthermore, fine-tuning – the process of adjusting and refining a pre-trained language model for a specific task or domain – allows for additional customization by exposing the model to a targeted dataset (Demszky et al. 2023; Zhong et al. 2023).

2.2.2 Empirical Evidence for LLM-Based Essay Scoring

In recent years, there has been a growing body of work focusing on the use of LLMs (and specifically ChatGPT) for AES (see, e.g., Mansour et al. 2024; Altamimi 2023; Mizumoto and Eguchi 2023; Latif and Zhai 2024). Numerous studies have demonstrated the effectiveness of ChatGPT for predicting categorical outcomes, such as classifying text as helpful/harmful (Touvron et al. 2023), identifying preferences (Lee et al. 2023), and detecting politeness/impoliteness (Ludwig et al. 2021). However, in studies focused on scoring continuous measures of writing quality or outcomes that require more nuanced evaluation, the LLM-based approach has been less successful (Ludwig et al. 2021; Mayfield and Black 2020).

In one recent study, Mizumoto and Eguchi (2023) used ChatGPT (text-davinci-003) to assess 12,100 essays, and found that the model demonstrated some alignment with human raters for classifying writing quality (low, medium, high), but showed limited differentiation between high- and low-quality essays. Similarly, Tate et al. (2024) compared ChatGPT with gold-standard human scoring, and found that ChatGPT exhibited greater internal consistency than human graders, though this was largely attributed to ChatGPT's tendency to assign scores in the mid-range rather than at the extremes. Research also emphasizes the impact of prompt design in this domain, with structured rubric-based prompts yielding more consistent results (Xia et al. 2024; Mansour et al. 2024). These findings collectively highlight the potential of LLMs, particularly ChatGPT, for AES, demonstrating their ability to achieve reasonable agreement with human raters and even surpass them in internal consistency. However, challenges remain in differentiating fine-grained quality levels and scoring essays at the extremes. Additionally, there is limited research directly comparing the efficacy of existing ML-based AES against the potential of ChatGPT.

2.3 Research Questions

In this study, we perform a comprehensive evaluation assessing the efficacy of ChatGPT for essay scoring under various prompt designs, fine-tuning choices, and outcome types, with the goal of replicating the gold-standard human grading as closely as possible. By comparing LLM-based and ML-based grading methods, our findings offer insights into the strengths and limitations of each approach for automated essay scoring. Specifically, we aim to address the following questions:

1. How do LLMs (e.g., ChatGPT) compare to established ML methods for predicting both continuous and categorical outcomes in the context of essay scoring?
2. Does the type of outcome impact the relative performance of LLM-based and ML-based scoring methods?

3 Data

In our empirical evaluation, we compare the performance of LLM-based and ML-based methods for scoring student-generated essays in two different contexts. First, we consider a collection of argumentative essays on iPad usage written by students in grades 4-7. Our second dataset consists of expository essays on the topic of censorship written by students in grade 10.

3.1 iPad Usage in Schools

We used data from the Catalyzing Comprehension through Discussion and Debate (CCDD) study, a randomized control trial that designed and implemented the Word Generation program (WordGen; Snow et al., 2009). From 2012 to 2014, the study examined the effect of the WordGen intervention to promote deep reading comprehension and academic language (see Jones et al., 2019). The WordGen study comprised two groups in grades four through seven: the treatment group, who engaged in discussions and debates about contestable topics from the Word Generation curriculum, followed by writing persuasive essays about the topics; and the control group, who did not participate in these activities. Additionally, students in both groups wrote persuasive essays about whether iPads should be used in their schools. For our analysis, we examine a sample of 2,687 of these essays sourced from 23 schools in four districts in the northeastern United States. Within our sample, essays ranged in length from 1 to 539 words, with an average essay length of 127.5 words. Even minimal student responses (e.g., "yes") were included in the analysis.

Essay Grading Rubric. A team of seven research assistants, experienced in English language teaching, scored and classified the essays. Training involved collaborative scoring of a set of six essays, guided by the holistic writing rubric (NAEP, of Educational Progress (NAEP) 2017), and facilitated by group discussions to reconcile any discrepancies. The rubric's evaluation criteria included:

1. Development of Ideas: Assessing the depth, complexity, richness of details, and examples in the essay.
2. Organization: Evaluating the logical flow of ideas through text structure, coherence, and focus.
3. Language Facility and Convention: Examining the clarity and effectiveness in sentence structure, word choice, voice, tone, grammar, usage, and mechanics.

Following this training phase, each research assistant independently scored 135 essays from the full sample. Subsequent to their independent assessments, the team compared their ratings to reach a consensus on the final scores. Inter-rater reliability, assessed using Kendall's Coefficient of Concordance for Ordinal Response, exceeded

0.7, indicating a marginally acceptable level of agreement among research assistants (Field 2005). In addition to evaluating writing quality, essays were categorized into five distinct "stances" regarding the use of iPads in schools: Affirmative (allow iPads in school), Negative (do not allow iPads in school), Balanced (allow iPads in school with restrictions), Ambivalent (not clear on stance) and, No Argument (not an argumentative stance and off-topic). For our analyses, we aggregated 'Balanced', 'Ambivalent', and 'No Argument' into a single 'Other' stance due to low baseline frequencies in these categories.

3.2 Censorship in the Libraries

To investigate the adaptability and robustness of our methodology across different writing prompts and student populations, we applied the same evaluation strategy in a different context using data from the Automated Student Assessment Prize (ASAP) corpus (Hamner et al. 2012), a large open-source data set of human-graded essays that has been extensively used in evaluations of AES systems (Lagakis and Demetriadis 2021).

In particular, we focus on a subset of the ASAP corpus consisting of persuasive essays written by 10th-grade students about censorship in libraries (prompt 2). This dataset comprises 1,800 essays, with lengths ranging from 33 to 1,149 words and an average length of 419.10 words. Comparing the results obtained from this dataset with those from the iPad usage essays allows us to assess the adaptability and robustness of our methodology across different topics and student populations.

Essay Grading Rubric. The essays in the dataset were graded for holistic writing quality on a scale of 1 (inadequate performance) to 6 (outstanding performance), considering four key domains:

1. Ideas and Content: This domain assesses the depth and breadth of the essay's ideas, ranging from a comprehensive exploration of the topic with rich details (6) to a failure to address the task with very few relevant ideas (1).
2. Organization: This domain evaluates the logical structure of the essay, considering the coherence and flow of ideas. Scores range from a well-organized presentation with smooth transitions (6) to a lack of organization with weak or absent transitions (1).
3. Style: This domain examines the writer's use of language, including vocabulary, sentence structure, and overall clarity. Scores range from exceptional word choice and varied sentence patterns (6) to minimal word usage, limited vocabulary, and problematic sentence patterns (1).
4. Voice: This domain assesses the writer's ability to express their unique perspective through appropriate language and tone. Scores range from an effective adjustment of language and tone with an original perspective (6) to inappropriate language and tone with a lack of originality (1).

Following this rubric, each essay was scored by a trained human rater. See Hamner et al. (2012) for additional details on the scoring process.

4 Methods

We compared various automated methods to score the essays, trying to replicate the gold-standard human coding. We have two general types of approach[1]: machine learning-based grading, and large language model-based grading. Details of each approach and their evaluation processes are described below.

4.1 ML-Based Scoring

4.1.1 Text Pre-processing and Feature Extraction

Prior to feature extraction, we performed standard text pre-processing steps including spelling correction, removal of punctuation and non-character symbols, and conversion of all text to lowercase. We then generated a diverse set of text features capturing various lexical, syntactic, and psychological aspects of the essays. These features ranged from basic frequency-based text statistics, such as word count and average word length, to more complex measures of writing style, including clout and emotional tone. For feature extraction, we employed several tools: the `rcttext` package in R (Mozer et al. 2024), Linguistic Inquiry Word Count (LIWC) software (Boyd et al. 2022), and the Tool for the Automated Analysis of Cohesion (TAACO) (Crossley et al. 2016). Initially, our statistical feature set included 315 indices for each dataset. However, to improve model performance, we refined this feature set by removing collinear features and those with near-zero variance. After this refinement process, the final feature sets used for training the machine learning models consisted of 265 features for the 'iPad usage in schools' dataset and 282 features for the 'censorship in libraries' dataset.

4.1.2 Evaluation Process

To derive ML-based predictions for both regression (writing quality scores) and classification (essay stance/categorized quality) tasks, we followed a systematic four-step approach:

1. Step 1: Choose a random sample of n essays to serve as the training data.
2. Step 2: Train ML models to predict the target outcome usingautomatically-extracted text features, with tuning parameters selected via cross-validation.
3. Step 3: Apply the trained models to predict scores/stances for the remaining documents.
4. Step 4: Compare ML predictions against human-coded scores/stances in the held-out sample.

We evaluated out-of-sample prediction performance across a range of training sample sizes varying from 5% to 90% of the full sample, to understand how prediction

[1] We also explored a third hybrid method using ChatGPT predictions as additional input features in our ML models. Following Mizumoto and Eguchi (2023), this approach aimed to assess if the two methods could complement each other to achieve an even better scoring model. However, this hybrid approach did not lead to meaningful changes in model performance for either dataset in either the regression or classification tasks.

accuracy varies with training set size. For model specification, we considered 4 different types of ML models: random forest (RF), regularized random forest (RRF), stochastic gradient boosting (GBM) and gradient extreme boosting (XGBOOST). Each of these methods can handle both regression and classification problems with minimal modification – for regression, they predict continuous values, while for classification, they predict class probabilities and assign the most likely category. All models were fit using functionality provided by the `caret` package (Kuhn 2015) in R. Our choice of tree-based methods aligns with the findings of Mozer and Miratrix (2025), who found that tree-based models (particularly random forests) outperformed other methods for predicting a continuous measure of writing quality.

For each training size, we generated 100 random test-train splits (reserving 80% and 20% of the data for training and testing, respectively). For each training sample, we used 10-fold cross-validation to tune model hyperparameters and select the configuration that yielded the best predictive performance based on the cross-validated metric (e.g., RMSE or classification accuracy). The final model, trained on the entire training set using the selected hyperparameters, was then applied to the held-out sample for evaluation.

To address potential issues arising from imbalanced category frequencies, we explored various stratified subsampling strategies. In classification tasks, imbalance in categories can lead to skewed predictions, where models disproportionately favor majority classes, reducing classification accuracy for underrepresented categories and potentially introducing bias (He and Garcia 2009). To mitigate this issue, we explored common subsampling techniques such as down-sampling (i.e., randomly reducing the number of instances from the majority class to match the minority class), up-sampling (i.e., increasing the number of instances in the minority class by duplicating them), and the Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al. 2002), which generates synthetic samples for the minority class based on the existing data, effectively increasing its size and diversity. However, simple random sampling, where each observation has an equal chance of being selected, consistently produced the best results across our evaluation metrics. Therefore, we opted to present simple random sampling for the final results.

4.2 Large Language Model-Based Grading

4.2.1 Prompt Variations

Prompt design is crucial for developing LLM-based scoring systems, as it directly influences model output (Wu et al. 2023). We adopted the prompt structures used in Lee et al. (2023), but refined them through pilot studies to identify the most consistent and reliable prompts for our specific task. This process led to the development of three different prompts, each providing increasing levels of guidance to the LLM:

1. Base Prompt: Essential instructions on essay scoring or classification criteria
2. Few-Shot Prompt: Builds on the base prompt by including a sample of graded essay examples, offering the LLM more context
3. Few-Shot + CoT Prompt: Enhances the few-shot prompt by incorporating Chain of Thought (CoT) reasoning for ideally improved alignment with human grading

In the context of LLMs, zero-shot learning refers to a technique where an LLM is prompted without any examples, attempting to leverage its inherent reasoning patterns (Liu et al. 2023). However, in this study, a pure zero-shot learning approach was not feasible as the prompts needed to include detailed instructions on how to classify student essays into three stances or assess them on a scale of 1-7. This comprehensive prompt, termed the 'base prompt,' serves as the foundation for all subsequent prompts used in the study. While the prompt structures for the iPad usage and censorship datasets were similar, the scoring rubrics differed significantly in length (46 words for iPad usage vs. 803 words for censorship). This discrepancy could influence how the LLM interprets and applies the rubric criteria, potentially affecting the consistency and accuracy of the final scores assigned by the LLM.

For the few-shot learning, we selected seven essays for quality scoring (one per score level) and five essays for essay stances (one per stance: Affirmative, Negative, Other - Ambivalent, Balanced, No Argument) from the iPad usage dataset. These were drawn from the essays used to train the research assistants (141 essays in total), and were distinct from the 2,687 essays used for subsequent grading. For the censorship dataset, we randomly selected 6 essays (one per score level) from the set of 1,800 essays. None of the selected essays for the few-shot approaches were used in subsequent grading.

In the few-shot+CoT condition of essay scoring, the iPad usage prompt included: "When evaluating and scoring the given text, consider three criteria (Development of Ideas, Organization, and Language Facility and Convention) and the examples above." Similarly, for essays with multiple stances, the prompt encouraged the few-shot+CoT approach: "If the essay incorporates multiple stances, determine the overall stance of the essay by considering the essay's coherence and intention and the examples provided above." For the censorship dataset, the few-shot+CoT prompt was: "When evaluating and scoring the given text, consider four domains (Ideas and Content, Organization, Style, and Voice) and the examples above." The investigation on these prompts demonstrates whether the Few-Shot and Few-Shot + CoT prompts enhance the performance of the LLM in aligning with human grading standards.

See Appendices A and B of the supplement for additional details on the final prompts used for each dataset and outcome measure.

4.2.2 Fine Tuning

We also fine-tuned the ChatGPT model using a selection of essay examples from our text data, to tailor ChatGPT for essay evaluation and replicate the human grading process. OpenAI (2024) recommends starting with at least 10 examples for fine-tuning, with 50-100 examples typically yielding noticeable improvements with GPT-3.5-turbo. However, they also emphasize that the optimal number depends on the specific use case.

Following this guidance, for the iPad usage dataset (2,687 essays), we randomly selected 30 essays per essay opinion (90 essays total) and 13 essays per quality score (91 essays total) to fine-tune ChatGPT. We then reserved five essays per essay opinion (15 essays total) and two essays per quality score (14 essays total) from the remaining pool for validating the correctness of the fine-tuned model. For the censorship dataset

(1,794 essays), we employed a different selection strategy due to the distribution of scores. Scores of 1 and 6 were less frequent compared to scores 2-5, resulting in fewer examples available for fine-tuning. We randomly chose 11 essays for score 1, 15 essays each for scores 2-5, and two essays for score 6, resulting in a total of 65 essays for fine-tuning. For validation, we selected one essay for score 1, two essays each for scores 2-5, and one essay for score 6, yielding a total of 10 essays.

It is important to note the role of validation sets in ChatGPT's fine-tuning process. While fine-tuning in ChatGPT is possible without a validation set, including one allows the tuning process to generate reports based on this set. These reports are used to track progress and assess the need for further fine-tuning. The essays used in fine-tuning were excluded from the subsequent evaluation of ChatGPT's performance.

4.2.3 Evaluation Process

Unlike ML-based scoring, LLM-based scoring has no need to extract text features. We can directly apply LLMs to score the raw text. To get LLM-based essay scores and classify essays based on their stances, we followed these steps:

1. Step 1: Design a prompt for scoring or classifying essays
2. Step 2: Employ the LLM to evaluate or classify essays based on the prepared prompt
3. Step 3: Compare LLM-generated scores or stances with human-assigned grades to assess accuracy

We applied steps 1-3 using ChatGPT 3.5-Turbo-0125 (offered through the OpenAI API) across three distinct prompts: the base prompt, a few-shot prompt, and a few-shot prompt with CoT reasoning. We also evaluated both non-fine-tuned and fine-tuned versions of ChatGPT. To ensure consistency and comparability, we maintained a temperature setting of 0 (less randomness) across all models. The temperature parameter controls the randomness of ChatGPT's responses, with lower values producing more deterministic and focused outputs.

4.3 Evaluation Metrics

To comprehensively assess and compare performance of ML and LLM-based scoring methods, we used standard metrics that are suitable for both regression and classification tasks, and are widely used in the AES literature (Ramesh and Sanampudi 2022).

For regression (i.e., assessing writing quality scores), we employed two commonly used performance metrics: Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). RMSE quantifies the average squared difference between model-generated ratings and human-rated scores, with lower values indicating more accurate predictions. Conversely, R^2 represents the proportion of variance in human-coded essay scores explained by model predictions, with higher values signifying better performance.

For classification tasks (i.e., essay stance and categorized writing quality), we used three metrics: accuracy, unweighted Kappa (UWK), and Quadratic Weighted

Kappa (QWK). Classification accuracy measures the proportion of correct predictions, while UWK and QWK measure agreement between model predictions and true (i.e. human-coded) labels beyond what is expected by chance alone. Following conventional standards (Landis and Koch 1977), we consider UWK values of 0.41-0.60 as moderate agreement, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect agreement. For QWK, which gives greater penalties to disagreements that are further apart on the ordinal scale (e.g., predicting "low" when the true label is "high" is penalized more heavily than predicting "medium"), values greater than or equal to 0.70 are considered acceptable (Williamson et al. 2012).

5 Results

Tables 1 and 2 show the performance of the standard and fine-tuned ChatGPT models as well as the tree-based machine learning algorithms for both the regression and classification tasks, using both the iPad usage and censorship datasets. Additionally, Figures 1–3 visually illustrate how the performance of the ML models varied with different proportions of training data, providing further insights into the relative strengths and weaknesses of these automated scoring methods.

5.1 Regression (Predicting Writing Quality)

Table 1 presents the performance of both the ML-based and LLM-based approaches for assessing writing quality on continuous scales (1-7 for iPad usage and 1-6 for censorship). The results reveal notable differences in performance across both datasets.

First, ChatGPT's performance in predicting writing quality varied considerably, with R^2 values ranging from 0.16 to 0.39 for the iPad usage dataset and from 0.23 to 0.46 for the censorship dataset. These values indicate generally poor success in predicting human-coded scores (the gold standard). In terms of error metrics, ChatGPT models yielded RMSE values between 1.12 and 2.00 for the iPad dataset and between 0.87 and 1.49 for the censorship dataset. Across all configurations, base prompts consistently outperformed both few-shot and few-shot + CoT approaches, and fine-tuning generally enhanced overall model performance, though the improvements were not always substantial.

In contrast, tree-based ML methods (fit using an 80% training and 20% test split) demonstrated considerably stronger performance in predicting writing quality. For the iPad usage dataset, ML models achieved R^2 values of 0.71–0.73 and RMSE values of 0.59–0.61, while for the censorship dataset, they yielded R^2 values of 0.57–0.59 and RMSE values of 0.50–0.51. Notably, the performance gap between ML and LLM-based approaches is evident in both performance metrics. For instance, for the censorship dataset, even the best-performing ChatGPT configuration (fine-tuned base prompt), which achieved an R^2 of 0.46, still had a substantially higher RMSE (0.87) compared to any of the ML models (roughly 0.50). This finding suggests that while LLMs may capture some patterns in essay quality, they lack the precision of feature-driven ML approaches for this specific task.

Figure 1 further illustrates how ML model performance improves with increasing percentages of training data. Interestingly, even with minimal training data – only

5% of the iPad usage data and 10% of the censorship data – the ML-based approach resulted in better performance than the best ChatGPT configurations in terms of both R^2 and RMSE. Differences between the four tree-based ML models themselves were not substantial, and their performance converged as the size of the training data increased, further suggesting that the advantage of this approach lies more in the use of text-derived features for prediction rather than the choice of a specific model.

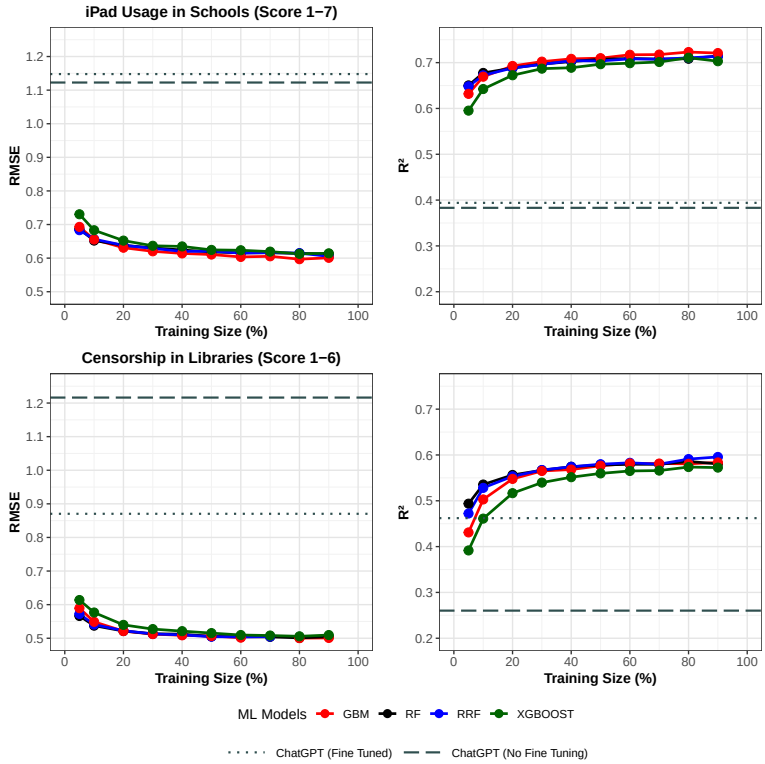

Fig. 1 Performance of ML models for predicting writing quality (regression) in terms of RMSE (left; lower values indicate better performance) and R^2 (right; higher values indicate better performance) with varying training data percentages compared to ChatGPT baselines (dashed lines). Top panel shows results for iPad usage dataset results and bottom panel shows censorship dataset results

Table 1 Comparison of ChatGPT and ML models for predicting writing quality (regression) and essay stance (classification). RMSE (Root Mean Squared Error) and R^2 measure predictive accuracy for writing quality scores; Accuracy, UWK (Unweighted Kappa), and QWK (Quadratic Weighted Kappa) assess classification performance for essay stance across iPad usage and censorship datasets.

	iPad Usage in Schools									Censorship in the Libraries			
	Score (1-7)				Opinion (Aff, Neg, Other)					Score (1-6)			
Models	RMSE	(SE)	R^2	(SE)	Acc.	UWK	(SE)	QWK	(SE)	RMSE	(SE)	R^2	(SE)
ChatGPT													
Base	1.12	(0.016)	0.38	(0.014)	0.82	0.65	(0.014)	0.44	(0.021)	1.22	(0.015)	0.26	(0.021)
Few-shot	1.66	(0.019)	0.17	(0.014)	0.84	0.68	(0.013)	0.50	(0.020)	1.49	(0.016)	0.24	(0.019)
Few-shot CoT	1.45	(0.018)	0.16	(0.014)	0.87	0.73	(0.012)	0.58	(0.021)	1.47	(0.015)	0.28	(0.020)
Fined-Tuned ChatGPT													
Base	1.15	(0.017)	0.39	(0.015)	0.88	0.73	(0.013)	0.56	(0.021)	0.87	(0.012)	0.46	(0.018)
Few-shot	2.00	(0.023)	0.27	(0.016)	0.80	0.60	(0.014)	0.40	(0.020)	0.96	(0.014)	0.38	(0.020)
Few-shot CoT	1.57	(0.020)	0.36	(0.015)	0.84	0.68	(0.014)	0.50	(0.021)	1.00	(0.015)	0.38	(0.020)
Tree-Based Machine Learning													
RF	0.61	(0.003)	0.71	(0.003)	0.80	0.44	(0.005)	0.35	(0.006)	0.50	(0.002)	0.59	(0.004)
RRF	0.61	(0.002)	0.71	(0.002)	0.79	0.44	(0.005)	0.35	(0.006)	0.50	(0.003)	0.59	(0.004)
GBM	0.60	(0.002)	0.72	(0.002)	0.81	0.51	(0.005)	0.42	(0.007)	0.50	(0.002)	0.58	(0.004)
XGBOOST	0.61	(0.003)	0.71	(0.003)	0.81	0.52	(0.004)	0.42	(0.005)	0.51	(0.003)	0.57	(0.005)

Machine learning results based on 80%/20% train-test split. RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

5.2 Classification (Classifying Essay Stance)

The classification of essays regarding their stance (Affirmative, Negative, and Other) in the iPad usage dataset also revealed important differences between ML and LLM-based approaches, as shown in Table 1. Overall, ChatGPT outperformed ML models in classification of essay stance across all evaluation metrics. Here, ChatGPT's accuracy ranged from 80% to 88%, while UWK values ranged from 0.60–0.73, and QWK values from 0.40–0.58. Fine-tuning improved performance marginally, with the fine-tuned model using the base prompt achieving the highest accuracy (88%) and UWK (0.73). Interestingly, while fine-tuning improved performance across all metrics for the base prompt, it actually reduced effectiveness for few-shot and few-shot+CoT approaches. This suggests that, in some classification tasks, simpler prompting strategies may be more amenable to fine-tuning.

Tree-based ML models achieved slightly lower accuracy and agreement scores than ChatGPT, with similar performance metrics across all four models. Classification accuracy under the ML-based approach ranged from 79% to 81%, with UWK and QWK values of 0.44–0.52 and 0.35–0.43, respectively. Here, gradient boosting methods (GBM and XGBoost) exhibited better agreement with human scores than random-forest based methods (RF and RRF), as evidenced by their higher UWK and QWK scores.

Figure 2 further underscores the advantage of boosting methods over random forest approaches across a range of training data sizes. While accuracy rates were similar across models, boosting methods consistently outperformed random forest methods in terms of UWK (0.29-0.52 vs. 0.23-0.46) and QWK (0.27-0.44 vs. 0.18-0.36) across all training set sizes. These results are consistent with prior research suggesting that boosting methods are better suited for handling classification tasks with imbalanced classes (Galar et al. 2011).

In summary, ChatGPT demonstrated superior performance in essay stance classification, surpassing ML models in accuracy and agreement with gold-standard human labels. While fine-tuning provided marginal benefits, the few-shot+CoT prompting approach proved particularly effective when used in the non-fine-tuned setting.

5.3 Classification (Classifying categorical writing quality)

Given the previous results indicating that ML models excel at predicting continuous writing quality scores, contrasted with ChatGPT's superior performance for categorizing essay stances, we also investigated this pattern extends to other classification tasks – specifically, when using ordinal categorized scores derived from continuous data. To examine this question, we converted the continuous writing quality scores into three ordinal categories representing "Low", "Medium" and "High" quality essays (defined as Low (1-2), Medium (3-5), and High (6-7) for the iPad usage dataset, and Low (1-2), Medium (3-4), and High (5-6) for the censorship dataset).

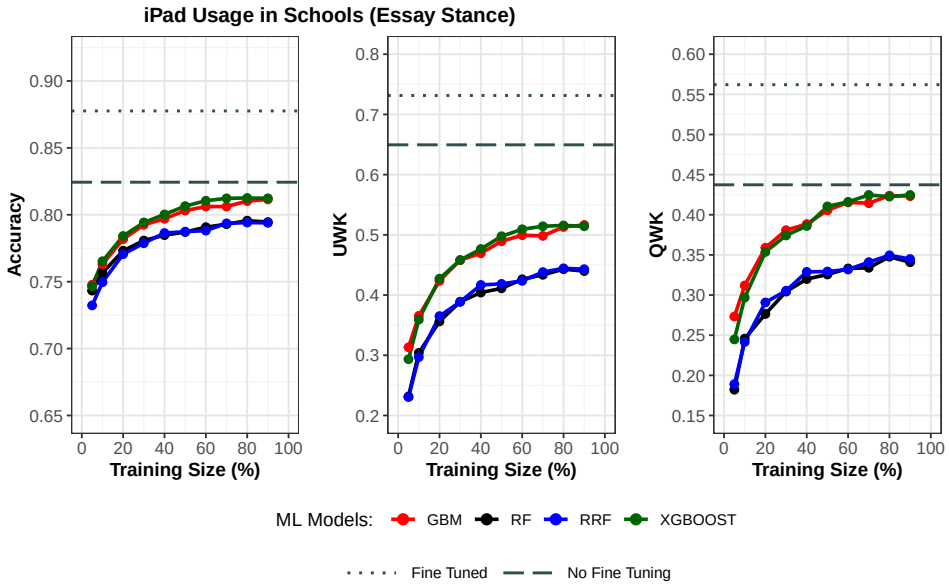

Fig. 2 Performance of ML models for classifying essay stance (Affirmative, Negative, Other) with varying training data percentages compared to ChatGPT baselines (dashed lines). Left: accuracy; Middle: unweighted kappa (UWK); Right: quadratic weighted kappa (QWK)

Table 2 Comparison of ChatGPT and ML models for predicting categorical writing quality scores (classification). Models are evaluated using accuracy, UWK, and QWK metrics to classify numeric writing quality scores into Low, Medium, and High categories for both datasets.

	Categorized Scores (1-Low, 2-Medium, 3-High)									
	iPad Usage in Schools					Censorship in the Libraries				
Models	Accuracy	UWK	(SE)	QWK	(SE)	Accuracy	UWK	(SE)	QWK	(SE)
ChatGPT										
Base	0.58	0.15	(0.013)	0.26	(0.012)	0.70	0.20	(0.022)	0.33	(0.021)
Few-shot	0.58	0.16	(0.014)	0.24	(0.017)	0.36	0.08	(0.011)	0.21	(0.016)
Few-shot CoT	0.66	0.18	(0.017)	0.23	(0.019)	0.66	0.23	(0.019)	0.32	(0.021)
Fined-Tuned ChatGPT										
Base	0.47	0.19	(0.011)	0.34	(0.013)	0.79	0.37	(0.026)	0.46	(0.024)
Few-shot	0.19	0.08	(0.006)	0.21	(0.012)	0.78	0.35	(0.025)	0.44	(0.023)
Few-shot CoT	0.19	0.08	(0.005)	0.21	(0.012)	0.79	0.35	(0.027)	0.45	(0.023)
Tree-Based Machine Learning										
RF	0.88	0.54	(0.005)	0.59	(0.005)	0.90	0.54	(0.008)	0.57	(0.008)
RRF	0.88	0.55	(0.007)	0.60	(0.006)	0.90	0.54	(0.010)	0.57	(0.009)
GBM	0.88	0.57	(0.006)	0.61	(0.006)	0.91	0.56	(0.008)	0.59	(0.007)
XGBOOST	0.88	0.56	(0.005)	0.61	(0.004)	0.91	0.55	(0.010)	0.58	(0.010)

Machine learning results based on 80%/20% train-test split. RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

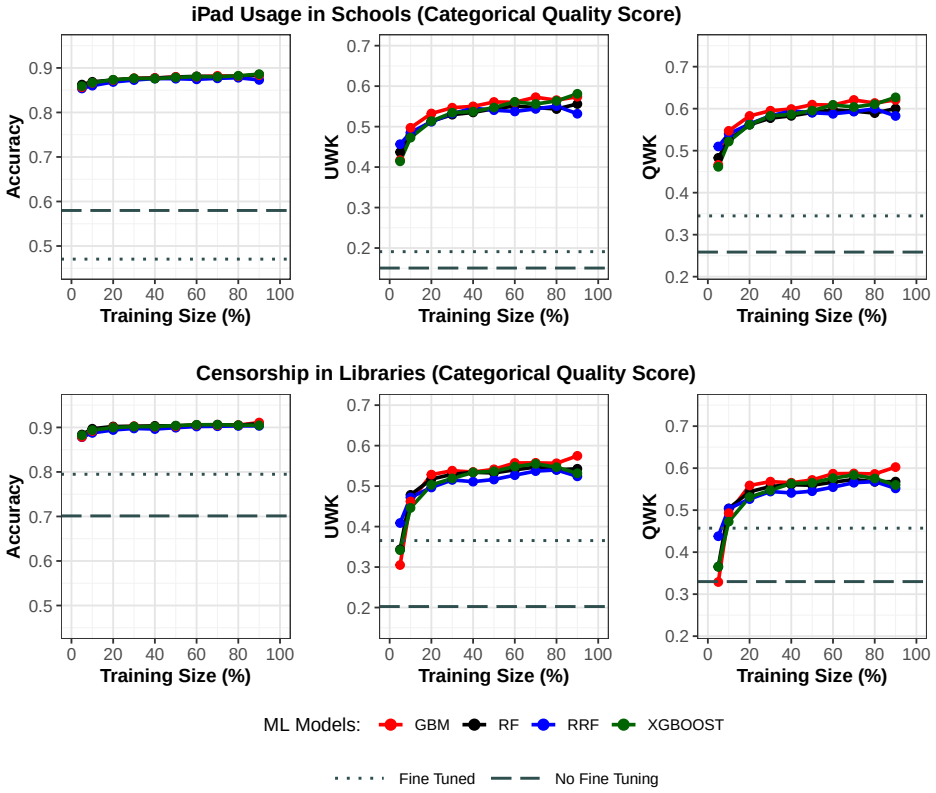

Fig. 3 Performance of ML models for classifying categorized writing quality scores (Low, Medium, High) with varying training data percentages compared to ChatGPT baselines (dashed lines). Top panel shows iPad usage dataset results and bottom panel shows censorship dataset results. Left: accuracy; Middle: unweighted kappa (UWK); Right: quadratic weighted kappa (QWK)

As shown in Table 2, the performance of ChatGPT in classifying categorical writing quality varied substantially across different configurations and datasets. For the iPad usage dataset, the non-fine-tuned model combined with the few-shot + CoT prompting approach demonstrated the highest classification accuracy (66%), but showed poor agreement with human-coded labels based on both the UWK (0.18) and QWK (0.23). Conversely, in the fine-tuned setting, the base prompt resulted in the highest accuracy (47%), with UWK (0.19) and QWK (0.34) values again falling within the unacceptable range. Interestingly, while all fine-tuned models showed worse performance than their non-fine-tuned counterparts, the fine-tuned base prompt still achieved the highest UWK and QWK among all prompts. For the censorship dataset, in the non-fine-tuned setting, the base prompt exhibited the highest accuracy (70%) and QWK (0.33). In

this case, unlike for the iPad usage dataset, fine-tuning substantially improved performance across all prompts, with all fine-tuned models showing comparable performance in terms of accuracy (78-79%), UWK (0.35-0.37), and QWK (0.44-0.46). These contrasting results between the two datasets underscore how performance patterns can vary considerably depending on the specific text and assessment task at hand.

For the iPad usage dataset, ML models achieved accuracy rates of 88%, UWK values of 0.54-0.57, and QWK values of 0.59-0.61. For the censorship dataset, accuracy rates reached 90-91%, with UWK values of 0.54-0.56 and QWK values of 0.57-0.59. All four ML approaches exhibited comparable performance, with minimal differences between random forest and boosting methods, unlike in the essay stance classification task where boosting methods showed clear advantages.

Figure 3 illustrates ML model performance under different training sample sizes for this task. While lower performance in terms of UWK and QWK was observed for the smallest training set sizes (up to 10%), ML models still consistently outperformed the best ChatGPT configurations across all metrics and for both datasets. Notably, all four ML models showed better performance for classifying categorized writing quality scores than for classifying essay stances. These results indicate that regardless of whether writing quality is measured on a continuous scale or categorized into ordinal groups, ML-based approaches consistently outperform LLM-based methods in evaluating writing quality. This advantage holds across both of our datasets, which represent different age groups, essay topics, and scoring rubrics, suggesting a fundamental difference in how feature-based ML models assess textual quality compared to LLMs.

6 Discussion

The current study explored the potential of both tree-based ML models and LLMs for AES. Our results offer several valuable insights into the strengths and limitations of these approaches and their suitability for different AES tasks.

6.1 Quality of Writing

Our analysis reveals interesting dynamics in the performance of tree-based ML methods and LLMs (specifically ChatGPT) for both regression (predicting continuous writing quality scores) and classification (categorizing writing quality levels).

For predicting writing quality, tree-based machine learning (ML) algorithms consistently outperform ChatGPT, regardless of whether the outcome variables are continuous or categorical, emphasizing that the nature of the outcome is more important than the outcome itself. The superior performance of tree-based ML methods in this domain could be attributed to their ability to process the wide range of stylometric and linguistic features that define writing proficiency. Our finding highlights the efficacy of these models in capturing the intricate patterns and nuances within textual data that are indicative of writing quality, despite the fact that these models not directly interpret the logical essence of essays as humans do (Mozer et al. 2024). In contrast, ChatGPT's strength seems to lie more in detecting the presence of certain aspects or understanding contextual nuances, rather than providing precise

quality estimations. This observation suggests that the specific construct being measured plays a more significant role in model performance than the type of outcome (continuous vs. categorical).

6.1.1 Regression Analysis for Writing Quality Prediction

Interestingly, when using ChatGPT for quality assessment, simple base prompts often perform similarly or even better than more complex prompting techniques involving essay examples or CoT (few-shot and few-shot+CoT). This contrasts with previous research suggesting that more examples or including CoT improve in-context learning and overall results (Wu et al. 2023). Potential overfitting or a mismatch between these techniques and the assessment task may be factors. Zhong et al. (2023); Agrawal et al. (2022) support this notion, demonstrating randomly sampled in-context examples can degrade output quality when they have low correlation with the test data.

Regarding fine-tuning, our analysis suggests that fine-tuning generally improves overall ChatGPT performance for essay scoring, showing the potential of LLMs to understand complex language patterns and generate relevant assessments (Mizumoto and Eguchi 2023). However, the improvements were not consistent; in some instances, fine-tuning led to only marginal gains, particularly with the iPad dataset (non-fine-tuned base R^2: 0.38, fine-tuned base R^2: 0.39). These results do not align with previous findings on the performance benefits of fine-tuning (Latif and Zhai 2024; Wei et al. 2021; Li et al. 2023). This inconsistency might stem from the use of randomly selected samples for fine-tuning, echoing the findings of Zhong et al. (2023); Agrawal et al. (2022).

These findings show the unpredictable nature of fine-tuning and in-text learning techniques, as well as their dependence on the specific textual data being evaluated. Further research is needed to better understand the factors influencing how LLMs evaluate writing quality and to develop more robust and reliable methods for utilizing these models in text grading.

6.1.2 Classification Analysis for Writing Quality Prediction

Within ChatGPT models, we observed mixed results. For non-fine-tuned models, base and few-shot results were inconsistent across datasets. The few-shot +CoT prompt provided the highest accuracy and UWK in the iPad usage dataset, but not the highest QWK. Conversely, in the censorship dataset, the base prompt achieved the highest accuracy and QWK, but not the highest UWK. Among fine-tuned ChatGPT results, the base prompt performed best in the iPad usage dataset, while all prompt performances were similar in the censorship dataset. Additionally, overall fine-tuned ChatGPT performed worse than non-fine-tuned ChatGPT in the iPad usage dataset, while the opposite was true in the censorship dataset. These mixed results across the two datasets indicate that performance patterns can vary depending on the specific textual data being evaluated.

6.1.3 Classification of Essay Stance

We found that LLMs generally outperform tree-based ML models in the categorical task of classifying essay opinions, a finding that aligns with previous studies on categorical outcomes classification (Lee et al. 2023; Touvron et al. 2023; Ludwig et al. 2021). This implies that LLMs, with their advanced understanding of contextual language nuances and ability to detect subtleties in text, may excel in tasks that require the interpretation of essay stances within written content.

Within ChatGPT models, specific techniques yielded interesting results. Without fine-tuning, few-shot+CoT prompts offered the best outcomes, supporting prior research on the benefits of additional prompting strategies (Li et al. 2023; Wu et al. 2023). However, with a fine-tuned ChatGPT, base prompts demonstrated the highest performance, suggesting that providing additional examples during fine-tuning may sometimes hinder output quality in opinion classification tasks. This further reinforces the potential issues discussed in the previous section regarding the unpredictable nature of fine-tuning and its dependence on specific text data.

6.2 Limitations and Future Directions

Our research provides valuable insights into the efficacy of ML- and LLM-based AES systems, contributing to the development of more efficient and scalable methods for assessing student writing. However, it is important to acknowledge limitations and potential areas for future research. First, we note that the performance of any LLM-based system can vary depending on the specific model and prompt used for evaluation. Future studies could investigate the performance of other open-source LLMs (e.g., Claude, Gemini, etc.) or explore the use of more advanced prompting strategies such as retrieval-augmented generation (RAG; Gao et al. 2023). Second, we note that this study focused on essays from students in grades 4-7 for the iPad usage dataset and grade 10 for the censorship data. Future research could incorporating essays from a wider range of grade levels, covering diverse topics and varying in quality, and utilizing different rubrics to enhance generalizability. Careful attention also should be given to the selection of in-context examples for prompting techniques (like few-shot) and training examples for fine-tuning. Mitigating the potential negative impacts of randomly chosen examples could provide clearer insights into the true effects of these strategies within the AES context.

6.3 Implications

Our findings have several practical implications for the development and implementation of AES solutions. Tree-based ML methods remain robust tools for assessing essay quality, generally exceeding, by large margin, LLMs in this domain. Conversely, LLMs demonstrate a particular strength in tasks like essay opinion classification. This highlights the importance of matching the AES approach best suited to the specific scoring scenario. Importantly, the success of base prompts underscores that overly complex prompting techniques may not always be necessary for effective AES. Moreover, using a common rubric for essay scoring in LLMs may not be effective, suggesting that tailored rubrics for LLMs are needed.

The unpredictable nature of fine-tuning LLMs warrants careful consideration. Practitioners should weigh the cost and potential benefits of fine-tuning, as fine-tuned outcomes do not guarantee improvement; results can vary based on the specific texts provided or the LLM version used. It is also crucial to note that the results obtained using the ChatGPT API may differ from those generated through the web interface, which has additional capabilities (e.g., memory) not mirrored in the API. Finally, the field of LLMs evolves rapidly, and while our results offer valuable insights into their current capabilities, researchers and practitioners should remain aware that the emergence of newer models could reshape the landscape of these AI-powered scoring methods.

Overall, while using ML- or LLM-based AES to fully replace human grading still remains a distant goal, these approaches hold great promise as powerful tools to augment the grading process. Their susceptibility to biases and less transparent decision-making processes require further research and development to ensure fairness and responsible implementation. For educators and assessment designers, these findings underscore the importance of aligning AI-based scoring with pedagogical priorities. While LLMs may streamline grading, their inconsistent alignment with human judgment suggests that automated feedback should complement, rather than replace, expert evaluation. Additionally, rubric design plays a crucial role in scoring accuracy, as LLMs struggle with complex or ambiguous scoring criteria. Assessment frameworks should be structured with AI's capabilities and limitations in mind, to ensure that scoring criteria are both interpretable for AI models and pedagogically meaningful for students and educators. These findings and insights provide helpful guidance for researchers, educators, and policy makers looking to improve the efficiency and consistency of automated essay assessments, while maintaining the essential role of human expertise and judgment.

Statements and Declarations

Funding

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant *R305D220032*. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Data and Code Availability

The code and data supporting the findings of this study are available at https://github.com/reaganmozer/ChatGPT-vs-ML-replication

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by YK and RM. The first draft of the manuscript was written by YK and all authors contributed to writing (review and editing) on previous versions of the manuscript. All authors read and approved the final manuscript.

References

Altamimi, A.B.: Effectiveness of chatgpt in essay autograding. In: 2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 102–106 (2023). IEEE

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., Ghazvininejad, M.: In-context examples selection for machine translation (2022) arXiv:2212.02437 [cs.CL]

Boyd, R.L., Ashokkumar, A., Seraj, S., Pennebaker, J.W.: The development and psychometric properties of liwc-22. Austin, TX: University of Texas at Austin **10**, 1–47 (2022)

Božić, V., Poola, I.: Chat GPT and Education vol. 10, (2023)

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002) https://doi.org/10.1613/jair.953

Crossley, S.A., Kyle, K., McNamara, D.S.: The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. Behavior research methods **48**, 1227–1237 (2016)

Dikli, S.: An overview of automated scoring of essays. The Journal of Technology, Learning and Assessment **5**(1) (2006)

Demszky, D., Yang, D., Yeager, D.S., Bryan, C.J., Clapper, M., Chandhok, S., Eichstaedt, J.C., Hecht, C., Jamieson, J., Johnson, M., *et al.*: Using large language models in psychology. Nature Reviews Psychology **2**(11), 688–701 (2023)

Dong, F., Zhang, Y., Yang, J.: Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 153–162 (2017)

Ekin, S.: Prompt engineering for chatgpt: a quick guide to techniques, tips, and best practices. Authorea Preprints (2023)

Field, A.P.: Kendall's coefficient of concordance. Encyclopedia of statistics in behavioral science (2005)

Foltz, P.W., Laham, D., Landauer, T.K.: The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning **1**(2), 939–944 (1999)

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(4), 463–484 (2011)

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 **2** (2023)

He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering **21**(9), 1263–1284 (2009)

Hussein, M.A., Hassan, H., Nassef, M.: Automated language essay scoring systems: A literature review. PeerJ Computer Science **5**, 208 (2019)

Heston, T.F., Khun, C.: Prompt engineering in medical education. International Medical Education **2**(3), 198–205 (2023)

Hamner, B., Morgan, J., lynnvandev, Shermis, M., Ark, T.V.: The Hewlett Foundation: Automated Essay Scoring (2012). https://kaggle.com/competitions/asap-aes

Kumar, V., Boulanger, D.: Explainable automated essay scoring: Deep learning really has pedagogical value. In: Frontiers in Education, vol. 5, p. 572367 (2020). Frontiers Media SA

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., *et al.*: Chatgpt: Jack of all trades, master of none. Information Fusion **99**, 101861 (2023)

Kuhn, M.: Caret: Classification and Regression Training

Lim, C.T., Bong, C.H., Wong, W.S., Lee, N.K.: A comprehensive review of automated essay scoring (aes) research and development. Pertanika Journal of Science and Technology **29**(3), 1875–1899 (2021)

Lagakis, P., Demetriadis, S.: Automated essay scoring: A review of the field. In: 2021 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–6 (2021). IEEE

Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics, 159–174 (1977)

Ludwig, S., Mayer, C., Hansen, C., Eilers, K., Brandt, S.: Automated essay scoring using transformer models. Psych **3**(4), 897–915 (2021)

Lee, H., Phatale, S., Mansoor, H., Lu, K.R., Mesnard, T., Ferret, J., Bishop, C., Hall, E., Carbune, V., Rastogi, A.: Rlaif: Scaling reinforcement learning from human feedback with ai feedback (2023)

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys **55**(9), 1–35 (2023)

Latif, E., Zhai, X.: Fine-tuning chatgpt for automatic scoring. Computers and Education: Artificial Intelligence **6**, 100210 (2024)

Li, J., Zhao, R., Yang, Y., He, Y., Gui, L.: Overprompt: enhancing chatgpt through efficient in-context learning. arXiv preprint arXiv:2305.14973 (2023)

Mansour, W., Albatarni, S., Eltanbouly, S., Elsayed, T.: Can large language models automatically score proficiency of written essays? (2024)

Mayfield, E., Black, A.W.: Should you fine-tune bert for automated essay scoring? In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 151–162 (2020)

Mizumoto, A., Eguchi, M.: Exploring the potential of using an ai language model for automated essay scoring. Research Methods in Applied Linguistics **2**(2), 100050 (2023)

Mozer, R., Miratrix, L.: More power to you: Using machine learning to augment human coding for more efficient inference in text-based randomized trials. The Annals of Applied Statistics **19**(1), 440–464 (2025)

Mozer, R., Miratrix, L., Relyea, J.E., Kim, J.S.: Combining human and automated scoring methods in experimental assessments of writing: A case study tutorial. Journal of Educational and Behavioral Statistics **49**(5), 780–816 (2024) https://doi.org/10.3102/10769986231207886 https://doi.org/10.3102/10769986231207886

Nazir, A., Wang, Z.: A comprehensive survey of chatgpt: advancements, applications, prospects, and challenges. Meta-radiology **1**(2), 100022 (2023)

Educational Progress (NAEP), N.A.: Writing Framework of the 2017 National Assessment of Educational Progress. National Assessment Governing Board, U.S. Department of Education (2017). https://www.nagb.gov/naep-subject-areas/writing/framework-archive/2017-writing-framework.html

OpenAI Accessed May 9, 2024 (2024). https://platform.openai.com/docs/guides/fine-tuning

Ozdemir, S.: Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs, (2023).

https://books.google.com/books?id=aDvVEAAAQBAJ

Page, E.B.: Computer grading of student prose, using modern concepts and software. The Journal of experimental education **62**(2), 127–142 (1994)

Page, E.B.: Project essay grade: Peg. In: Shermis, M.D., Burstein, J. (eds.) Automated Essay Scoring: A Cross-disciplinary Perspective, pp. 43–54 (2003)

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations (2018). https://arxiv.org/abs/1802.05365

Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review **55**(3), 2495–2527 (2022)

Ramineni, C., Williamson, D.M.: Automated essay scoring: Psychometric guidelines and practices. Assessing Writing **18**(1), 25–39 (2013)

Sevcikova, B.L.: Human versus automated essay scoring: A critical review. Arab World English Journal **9**(2) (2018)

Shermis, M.D., Hamner, B.: Contrasting state-of-the-art automated scoring of essays. In: Handbook of Automated Essay Evaluation, pp. 313–346 (2013)

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1882–1891 (2016)

Tate, T.P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., Warschauer, M.: Can ai provide useful holistic essay scoring? Computers and Education: Artificial Intelligence **7**, 100255 (2024)

Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., Tang, Y.: A brief overview of chatgpt: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica **10**(5), 1122–1136 (2023)

Weegar, R., Idestam-Almquist, P.: Reducing workload in short answer grading using machine learning. International Journal of Artificial Intelligence in Education **34**(2), 247–273 (2024)

Williamson, D.M., Xi, X., Breyer, F.J.: A framework for evaluation and use of automated scoring. Educational measurement: issues and practice **31**(1), 2–13 (2012)

Xia, W., Mao, S., Zheng, C.: Empirical study of large language models as automated essay scoring tools in english composition_taking toefl independent writing task for example (2024)

Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D.: Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert (2023)