

Literature Review

To determine the effectiveness of using LLMs for rating writing assessments, it is helpful to understand what has been done in the field. To that end, I will review to the history of Automated Essay Scoring. (e.g. Project Essay Grade, Latent Semantics Analysis, Neutral Networks, Pretrained transformer methods, and Large Language Models)

Automated Essay Scoring (AES)

Automated essay scoring (AES) is a technology that uses machines to grade and provide feedback on written work. The introduction of AES has been considered a solution to the high costs of writing assessment in addition to human raters' inconsistency in grading (Dikli, 2006; Hussein et al., 2019; Kumar et al., 2024; Lim et al., 2021; Link & Koltovskaia, 2023; Mizumoto & Eguchi, 2023). AES tools have been used in different testing contexts, including low-stakes assessments happening in classrooms and high-stakes assessments (Dikli, 2006), because of their advancements. AES has been improved over time, from using simple modals to predict scores based on surface features of written works, such as grammar and length, to using Large Language Models (LLMs), which can capture deeper elements in writing, such as semantics, pragmatics, and discourse.

Project Essay Grader® (PGE)

The earliest AES tool can be traced to Project Essay Grader® (PEG), invented by Dr. Ellis Page in 1966. PEG predicted scores by predicting "trins," which refers to elements like sentence structure, grammar, punctuation, and "proxies" such as text length, word length, and spelling. The tool used a statistical approach to predict scores based on the assumption that the quality of essays relies on surface elements like spelling, grammar, and text length. This tool did

not use natural language processing or take into account semantic aspects of writing (Valenti, Neri, & Cucchiarelli, 2003).

Whalen (1971) conducted a study using PEG to test if writing features used by Page were good predictors for writing quality written by seventh-grade students, and if the tool could predict writing scores in standardized English mechanics. The result showed that linguistics features used by Page(1966) were useful in the context. Multiple correlations were strong .85 for essay grade, .861 for mechanics, and .882 for language ability before correction. In addition, Whalen concluded that automatic essay evaluation achieved good agreement with human-raters in the standardized English mechanics tests despite the fact that human raters slightly outperformed the computer program.

PEG received support from some researchers. In a paper written by Daigon (1966), he claimed that tools like PEG can support teachers who do not have enough time for instructing composition in writing classes by grading. He stated that instead of having teachers spend time on counting clauses and grammar mistakes, computers can help them with that.

In 1997, Page et al. (1997) used PEG to study whether it can score essays based on traits, such as content, organization, style, mechanics, and creativity. When comparing the correlations of PGE with one human rater and those of one human rater with another human rater, PEG had higher correlations across the five traits. For example, PEG's correlation with an individual human rater on content was 0.68 while human rater's correlation with another human rater was 0.52. In this paper, Page encouraged the computerized scoring system's application in assessment because of its effectiveness and reliability.

Even though PEG could reach high agreement with human raters in some studies, it faced criticism for the ignorance of meaning aspects in writing. Chung and O'Neil (1997) claimed that

PEG simply took into account surface features of written essays and ignored meaning and content. Besides, since PEG relies heavily on examples and training sets, it requires a new dataset for each new set of essays. Lastly, the system can only score an essay based on the performance of other essays from the same essay population.

Latent Semantics Analysis (LSA)

Landauer and Dumais (1997) introduced LSA as a method that they believed could help explain Plato's problem, which concerns how people learn as much as they do from the limited information they receive. In their paper, LSA was described as a computational model that used general mathematical learning methods to acquire vocabulary knowledge from large bodies of English text at a rate comparable to schoolchildren. After experimenting with its performance, they concluded that LSA learns a great number of word-meaning similarities, roughly comparable to the knowledge of moderately competent English readers. They also found that about three quarters of LSA's vocabulary gain from reading came through indirect induction, meaning that the model learned about words even when those words did not appear directly in the passage. In addition, LSA could be used not only to learn relationships among individual words but also to represent aspects of text comprehension. This became an important foundation for understanding semantics and contextual knowledge in computational models.

Even though LSA was not created merely for assessment purposes, its application in the field was meaningful. Foltz et al. (1999) reviewed the effectiveness of the Intelligent Essay Assessor (IEA), which is an AES tool using LSA. His article showed that when comparing the essay against all previously graded essays, LSA's reliability in scoring was equivalent to that of human raters. Besides, the authors also suggested that IEA using LSA can be applied in different

contexts, such as it could be a consistency checker for teachers, or it can be used to help students improve their writing through automatic grades and comments.

Despite high correlation with human raters, LSA faced some criticism. Chung and O'Neil (1997) claimed that LSA would take enormous computational power due to the use of matrices and vectors to represent data and to determine similarities among texts in addition to a huge source of corpus used for training data. Miller (2003) also agreed with this point in an article reviewing the application of LSA in writing assessment. In addition, Miller believed that the reason for LSA to reach that high correlation lies in the similar rating tendency between human raters and LSA. That is similar to human raters who tend to put more emphasis on content than on style or mechanism; LSA focuses more on semantics and ignores syntax and morphology.

Neural Networks

Neural networks were inspired by human brains, where sophisticated knowledge is learned by hundreds of billions of interconnected neurons. Researchers successfully demonstrated this intelligence on silicon (Wang, 2003). In the AES field, neural networks were highly rated by many researchers. They are believed to surpass traditional methods using predictive features, such as Support Vector Regression (SVR) or Bayesian Linear Ridge Regression (BLRR) (Alikaniotis et al., 2016; Dong & Zhang, 2016; Taghipour & Ng, 2016). Alikaniotis et al. (2016) compared different neural network models to traditional baselines. In this study, they combined different word embedding techniques (word2vec and score-specific word embedding) and different neural network models, such as Long Short-Term Memory and Bidirectional Long Short-Term Memory. Then they compared these models with traditional baselines, including Support Vector Regression models and doc2vec. The results showed that their best model was SSWE + two-layer bidirectional LSTM. It achieved the strongest

performance in their table: Spearman's $\rho = .91$, Pearson's $r = .96$, RMSE = 2.4, and Cohen's kappa = .96. This outperformed the SVM baseline and the other neural models.

A similar result was gained in another study conducted by Dong and Zhang (2016). In this paper, they used convolutional neural networks (CNN) without using any specific word embedding techniques and traditional baselines, including Bayesian Linear Ridge Regression (BLRR) and Support Vector Regression (SVR). These models were used to grade essays based on 8 prompts of different genres, and they are tested in both an in-domain environment (the model is trained and tested on the same type of data) and a cross-domain environment (the model is trained on one type of data but tested on a different type of data). For in-domain, CNN appeared to perform significantly better than baseline model SVR, while it is competitive to BLRR. For cross-domain, since traditional baselines use features that are not the best for cross-domain tasks, the authors used ML-p to support them in performing cross-domain tasks. The result showed that CNN outperformed ML-p on almost all pairs of adaptation experiments. The authors concluded that automatically learned features used by neural network methods are competitive compared to handcrafted features used by traditional baselines, which makes deep learning become potentially valuable in AES.

In 2019, Cai (2019) combined the use of neural models with selective handcrafted features and a word embedding technique (GloVe) to find out if they improve rating performance. The features are at surface levels such as spelling, essay length, and vocabulary complexity. The results showed that adding manual features might lead to improvement, but it is not significant. Besides, the use of Glove in embedding words affected the performance significantly.

Pretrained transformer methods

After the neural networks area, pretrained transformer methods such as BERT and RoBERTa were considered the most state-of-the-art methods in natural language processing. Devlin et al. (2019) argued that previous language models were limited by their unidirectional design, which restricted the architectures available for pre-training. This was especially problematic for tasks that require context from both directions, such as question answering. Rodriguez et al. (2019) also conducted a study comparing two transformer methods, including BERT and XLNet, with older methods such as LSTM and Bag of Words (BOW). They used data from the Kaggle Automated Student Assessment Prize dataset and had these models predict essay scores and evaluated their performance mainly with quadratic weighted kappa (QWK). The result showed that individual BERT, XLNet, and LSTM models performed similarly. BERT had an average QWK of 74.75, XLNet had 74.34, and LSTM had 74.63, which suggests that the essay scoring problem has reached its ceiling in terms of modeling.

Furthermore, BERT was found to perform impressively in distinguishing registers in writing tasks (Ludwig et al., 2021). In this study, Ludwig and colleagues used BERT and BOW methods in discriminating if email responses were polite or impolite, noting that these emails were written in German. The results showed that the transform models could reach 0.92 accuracy to 0.94 accuracy, while the logistic regression model using BOW reached 0.84 accuracy. They suggested that transform models could be beneficial for reasoning tasks where the order of the words is of particular importance.

Large Language Models(LLMs)

Building on a type of neural network architecture called a transformer, large language models are trained on an immense amount of data that allows them to understand and generate

natural language (Stryker, n.d.). In language assessment, LLMs have received great attention from researchers, especially in writing assessment. The earliest studies are from 2023, after the introduction of ChatGPT. For example, Geçkin, Kızıldaş, and Çınar (2023) compared ChatGPT-3.5 scores with scores from five human raters for second-language academic writing. They introduced rubrics to ChatGPT-3.5 through the chatbox and to human raters. The results showed that scores assigned by human raters ranged from low to high positive correlation in statistical significance, and those by ChatGPT-3.5 ranged from slight to fair but were significant in agreement with two of the human raters. This suggests that ChatGPT-3.5 can be used as a useful writing assessment tool. Besides, scores generated by ChatGPT-3.5 were proved to be consistent in a study conducted by Mizumoto and Eguchi (2023). In this study, they used the TOEFL11 corpus, which contains 12,100 essays written by non-native English writers to grade, and the IELTS TASK 2 writing band descriptors (public version) as a rubric. Those 12,100 essays were graded first by ChatGPT-3.5, and after that the researchers used stratified sampling techniques to get 1,210 essays for re-scoring to assess the consistency of the initial scoring. The results showed that ChatGPT-3.5 achieved a certain level of accuracy and reliability for automated essay scoring. The results suggested that GPT could be useful as a support tool for writing evaluation, especially because it can score large numbers of essays quickly.

In addition to ChatGPT, other AES tools using LLM were studied to compare their performance with each other and with human raters. Pack, Barrett, and Escalante (2024) compared four major LLMs: Google's PaLM 2, Anthropic's Claude 2, OpenAI's GPT-3.5, and OpenAI's GPT-4, and they wanted to see how well these models could score essays compared with a pair of human raters and whether the models gave consistent scores over time. Using 119 essays from an English language placement test, each AES was used to score these essays twice

on two separate occasions. The results showed that the best-performing tool was ChatGPT-4, which demonstrated excellent intrarater reliability and good validity. In addition, all models improved their intrarater reliability over time except for GPT-3.5. Besides, GPT-4 correlated highest with human ratings; however, most correlations were moderate to strong.

In other two studies carried by Shermis (2024), ChatGPT-4o was used to rate two sets of essays, including ASAP essays and short-answer responses, and both were obtained from high-stakes assessments. In the two studies, the data sets were divided into test sets and sample sets. For sample sets, there were split into three proportions: 60% training sample, 20% test sample, and 20% validation sample. The procedures for the two studies started by analyzing the datasets separately, then the model was fed with the prompt, the source material, and scoring rubric. After that, it was given the training data, followed by the test dataset with only responses and no score information. Finally, ChatGPT was asked to give scores to those essays using four prediction models: linear regression, random forest, gradient boosting, and XGBoost. In Study 1, the result showed that ChatGPT had lower QWK scores than human raters in five of the nine datasets, but it matched or exceeded human agreement in four. For example, in Dataset 2a, human agreement was .80, while ChatGPT's highest QWK was .63. In Dataset 5, however, ChatGPT's gradient-boosting model scored .80, compared with .74 for human raters. In study 2, ChatGPT had lower QWK scores than human raters across all ten datasets. Although ChatGPT reached QWK values in the .70s, human agreement was often in the high .80s or .90s. Its best result was in Dataset 6, where the gradient-boosting model achieved .81, compared with .93 for human raters.

Furthermore, researchers started to pay more attention to different techniques to improve the performance of LLM tools. Lee et al. (2024) used three different prompt techniques in his study, including chain-of-thought prompting, zero-shot, and few-shot prompting to investigate

the effectiveness of ChatGPT using these techniques. The study showed that few-shot prompting worked better than zero-shot prompting. In simple terms, GPT scored more accurately when it was given examples of how responses should be scored. The authors reported that few-shot prompting reached about .67 accuracy, while zero-shot prompting reached about .60 accuracy. Besides, the results for chain-of-thought prompting showed that these tools worked better with a combination of question stem and rubric.

Mansour et al. (2024) examined the ability of ChatGPT and Llama 2 to score 12,978 essays from the ASAP dataset. The researchers used four prompts that gradually included additional information, such as the scoring rubric and a scored example, and evaluated the models using QWK. The results showed that prompt effectiveness depended on the model and task type. ChatGPT was more consistent than Llama across different prompts, but both models performed substantially below existing automated essay-scoring models. ChatGPT and Llama obtained average QWK values of .313 and .297, respectively, compared with .817 and .702 for the state-of-the-art models. Although the LLMs were not considered reliable for predicting essay scores, they showed potential for providing feedback to help students improve their writing.

Similarly, Song et al. (2024) compared three open-source large language models—Baichuan-13B, InternLM-7B, and ChatGLM2-6B—with statistical and deep-learning automated essay-scoring baselines. Quadratic weighted kappa was used to evaluate the agreement between the scores predicted by the models and the reference scores assigned by human raters. In the zero-shot condition, the open-source LLMs performed less effectively than the strongest baseline models; the best-performing open-source model, ChatGLM2-6B, achieved a QWK of .306, whereas the BERT-based Bi-LSTM baseline achieved a QWK of .660. The researchers then selected ChatGLM2-6B for additional few-shot and P-tuning experiments and found that its

performance improved to a QWK of .531 when it received six examples containing both human scores and teachers' comments. However, the 12-shot condition performed less effectively than the six-shot condition, which the authors suggested may have resulted from a model-forgetting problem caused by excessive context length. This finding is relevant to the present study because it suggests that the effectiveness of model adaptation may depend not only on the number of examples provided but also on the quality, structure, and method through which those examples are presented.

In addition, Shibata and Miyamura (2025) proposed a new method called LLM-based Comparative Essay Scoring (LCES). This new method asks the model to compare two essays and decide which one is better instead of having the models give essays a score. First, LCES will compare essays in pairs and turn them into ranking scores, which will be converted into final scores. Those final scores were compared with those generated by the LLM model using the zero-shot scoring method. The results show that LCES performs better than conventional zero-shot scoring methods while still keeping the process computationally efficient. The authors also find that LCES works across different LLM backbones, which suggests that the method is not limited to only one model.

Gap in the Literature

Although previous studies have examined different methods for adapting LLMs to automated essay scoring, few studies have directly compared supervised fine-tuning with prompt- and example-based customization. Some studies have investigated zero-shot, few-shot, P-tuning, or other prompting methods, but they have not compared a supervised fine-tuned Gemini model with a custom ChatGPT – 5.5 model within the same writing-assessment context. In addition, limited research has compared the costs and practical requirements of using these

two approaches for writing assessment. Therefore, the present study contributes to the literature by examining two practical options for adapting LLMs as automated writing-assessment tools and comparing their scoring performance, cost, accessibility, and scalability. This comparison may help institutions determine which approach is more suitable for scoring writing with an institution-specific proficiency rubric.