Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs

Changrong Xiao
Tsinghua University
Beijing, China
xcr21@mails.tsinghua.edu.cn

Wenxing Ma
Tsinghua University
Beijing, China
mawx21@mails.tsinghua.edu.cn

Qingping Song
Hong Kong Baptist University
Hong Kong, Hong Kong
qpsong@hkbu.edu.hk

Sean Xin Xu
Tsinghua University
Beijing, China
xuxin@sem.tsinghua.edu.cn

Kunpeng Zhang
University of Maryland
College Park, USA
kpzhang@umd.edu

Yufang Wang
Beijing Xicheng Educational Research Institute
Beijing, China
wangwang7587@163.com

Qi Fu
Beijing Xicheng Educational Research Institute
Beijing, China
lilyhoneypot@163.com

Abstract

Receiving timely and personalized feedback is essential for second-language learners, especially when human instructors are unavailable. This study explores the effectiveness of Large Language Models (LLMs), including both proprietary and open-source models, for Automated Essay Scoring (AES). Through extensive experiments with public and private datasets, we find that while LLMs do not surpass conventional state-of-the-art (SOTA) grading models in performance, they exhibit notable consistency, generalizability, and explainability. We propose an open-source LLM-based AES system, inspired by the dual-process theory. Our system offers accurate grading and high-quality feedback, at least comparable to that of fine-tuned proprietary LLMs, in addition to its ability to alleviate misgrading. Furthermore, we conduct human-AI co-grading experiments with both novice and expert graders. We find that our system not only automates the grading process but also enhances the performance and efficiency of human graders, particularly for essays where the model has lower confidence. These results highlight the potential of LLMs to facilitate effective human-AI collaboration in the educational context, potentially transforming learning experiences through AI-generated feedback.

CCS Concepts

• **Applied computing → Education**; • **Computing methodologies → Natural language processing**.

Keywords

LLM Application, Automatic Essay Scoring, AI-assisted Learning

ACM Reference Format:
Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3706468.3706507

1 Introduction

Writing practice is a crucial aspect of second-language learning, yet providing timely and reliable feedback remains a significant challenge for educators in China due to the high student-to-teacher ratio. This limitation often hinders students' academic progress, particularly for those who are motivated to pursue self-directed learning. Automated Essay Scoring (AES) systems offer a practical solution by delivering immediate and consistent feedback on students' work. These systems not only simplify the grading process for educators but also provide an efficient tool for learning analytics (LA), supporting both teaching and learning outcomes.

However, implementing AES systems effectively in real-world educational scenarios presents several challenges. First, the diverse range of exercise contexts and the inherent ambiguity in scoring rubrics complicate the ability of traditional models to deliver accurate scores. Second, interviews with high school teachers indicate that despite receiving accurate score predictions, they must still review essays to mitigate potential errors from the models. Consequently, relying exclusively on this system without human supervision is impractical in real-world scenarios. Thus, there is a clear need for AES systems that not only predict scores accurately but also facilitate effective human-AI collaboration. This should be supported by natural language explanations and additional assistive features to enhance usability.

This work is licensed under a Creative Commons Attribution International 4.0 License.

LAK 2025, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706507

To effectively tackle these challenges, it is crucial to highlight the latest advancements in the field of Natural Language Processing (NLP), particularly focusing on the development of large language models (LLMs). LLMs, such as OpenAI's ChatGPT [1], not only showcase impressive capabilities of robust logical reasoning but also exhibit a remarkable ability to comprehend and faithfully follow human instructions [25]. Furthermore, recent studies have highlighted the potential of leveraging LLMs in AES tasks [23, 24, 43].

In this study, we explore the potential of proprietary and open-source LLMs such as GPT-3.5, GPT-4, and LLaMA3 for AES tasks. We conducted extensive experiments with public essay-scoring datasets as well as a private collection of student essays to assess the zero-shot and few-shot performance of these models. Additionally, we enhanced their effectiveness through supervised fine-tuning (SFT). Drawing inspiration from the dual-process theory, we developed an AES system based on LLaMA3 that matches the grading accuracy and feedback quality of fine-tuned LLaMA3. Our human-LLM co-grading experiment further revealed that this system significantly improves the performance and efficiency of both novice and expert graders, offering valuable insights into the educational impacts and potential for effective human-AI collaboration. Overall, our study contributes three major advancements to the field:

- We pioneer the exploration of LLMs' capabilities as AES systems, especially in complex scenarios featuring tailored grading criteria. Leveraging dual-process theory, our novel AES framework demonstrates remarkable accuracy, efficiency, and explainability.
- We introduce an extensive essay-scoring dataset, which includes 13,372 essays written by Chinese high school students. These essays are evaluated with multi-dimensional scores by expert educators. This dataset significantly enhances the resources available for AI in Education (AIEd) and Learning Analytics (LA)[2].
- Our findings from the human-LLM co-grading task highlight the potential of LLM-generated feedback to elevate the proficiency of individuals with limited domain expertise to a level akin to that of experts. Additionally, it enhances the efficiency and robustness of human graders by integrating model confidence scores and explanations. These insights set the stage for future investigation into human-AI collaboration and AI-assisted learning within educational contexts.

2 Related Work

2.1 Automated Essay Scoring (AES)

Traditional Methods. Automated Essay Scoring (AES) stands as a pivotal research area at the intersection of NLP and education. Traditional AES methods usually involve a two-stage process, as outlined in [28]. First, features are extracted from the essay, including statistical features [22, 29] and latent vector representations [21, 26]. Subsequently, regression-based or classification-based machine learning models are employed to predict the essay's score [19, 20, 31, 35].

With the advancement of deep learning, AES has witnessed the integration of advanced techniques such as convolutional neural networks (CNNs) [6], long short-term memory networks (LSTMs) [36], and also pre-trained language models [18, 30]. These innovations have led to more precise score predictions, and state-of-the-art methods are primarily based on Bidirectional Encoder Representations from Transformers (BERT) [4, 39, 44].

The domain of AES has also experienced advancements with the incorporation of pre-trained language models to enhance performance. Lun et al. [18], Rodriguez et al. [30] utilized Bidirectional Encoder Representations from Transformers (BERT [5]) to automatically evaluate essays and short answers. Additionally, Yang et al. [44] improved BERT's performance by fine-tuning it through a combination of regression and ranking loss, while Wang et al. [39] employed BERT for jointly learning multi-scale essay representations.

LLM Applications in AES. Recent studies have explored The potential of leveraging the capabilities of modern LLMs in AES tasks. Mizumoto and Eguchi [23] provided ChatGPT with specific IELTS scoring rubrics for essay evaluation but found limited improvements when incorporating GPT scores into the regression model. In a different approach, Yancey et al. [43] used GPT-4's few-shot capabilities to predict Common European Framework of Reference for Languages (CEFR) levels for short essays written by second-language learners. However, the Quadratic Weighted Kappa (QWK) scores still did not surpass those achieved by the XGBoost baseline model or human annotators. Similarly, Han et al. [9], Stahl et al. [34] introduced prompting frameworks that did not outperform the conventional baselines.

Building on these insights, our study aims to further investigate the effectiveness of LLMs in AES tasks. We focus on more complex contexts and leverage domain-specific datasets to fine-tune LLMs for enhanced prediction performance. This research area offers promising avenues for future exploration and improvement.

2.2 AI-Assisted Decision Making

Researchers have extensively investigated human-AI teams, in which AI supports the decision-making process by providing recommendations or suggestions, while the human remains responsible for the final decision [38]. The objective of such human-AI collaboration is to achieve complementary performance, where the combined team performance exceeds that of either party operating independently [1]. To realize this, it is crucial to design an AI-assisted decision-making process that allows humans to effectively monitor and counteract any unpredictable or undesirable behavior exhibited by AI models [7]. This design aims to leverage the strengths of both humans and AI to enhance overall performance [10]. To our knowledge, no studies have yet investigated AES systems from this angle of collaborative co-grading.

2.3 Dual-Process Theory

Recent studies have developed architectures that imitate human cognitive processes to enhance the capabilities of LLMs, particularly in reasoning and planning [3]. According to dual-process theory in psychology [12, 41], human cognition operates via two distinct systems: System 1 involves rapid, intuitive "Fast Thinking",

---

[1] https://chat.openai.com

[2] Codes and resources are available in https://github.com/Xiaochr/LLM-AES

while System 2 entails conscious and deliberate "Slow Thinking" processes. LLM architectures inspired by this theory have been implemented in complex interactive tasks [17, 37], aiming to mitigate issues like social biases [13] and hallucination [2]. These adaptations have demonstrated improved performances in various areas.

3 Data

ASAP dataset. Automated Student Assessment Prize (ASAP[3]) dataset, stands as one of the most commonly used publicly accessible resources Automated Essay Scoring (AES) tasks. This comprehensive dataset comprises a total of 12, 978 essays, encompassing responses to 8 distinct prompts. Each essay has been evaluated and scored by human annotators. Essay sets are also accompanied by detailed scoring rubrics, each tailored with unique scoring guidelines and score ranges. These intricacies are essential as they cater to the multifaceted requirements and diverse scenarios of AES. The details of the ASAP dataset are presented in Table 1.

Our Chinese Student English Essay (CSEE) dataset. We have developed a novel English essay scoring dataset[4] specifically designed for AES tasks. The dataset was carefully curated in collaboration with 29 high schools in China, encompassing a total of 13, 372 student essays responding to two distinct prompts used in final exams. The evaluation of these essays was carried out by highly experienced English teachers following the scoring guidelines of the Chinese National College Entrance Examination (Table 8). Scoring was comprehensively assessed across three critical dimensions: Content, Language, and Structure, with an Overall Score ranging from 0 to 20. The descriptive statistics of this dataset are outlined in Table 2.

To protect privacy and address ethical considerations, all personal information regarding the students has been anonymized. Furthermore, consent for using the data exclusively for research purposes has been obtained from both teachers and students. Our CSEE dataset, derived from authentic English learning scenarios among Chinese students, represents a highly valuable and genuine resource for not only AES but also broader educational research.

4 Methods

In this section, we present the details of the models used in this study, including traditional AES baselines, LLM-based approaches, and our proposed Fast and Slow Thinking AES framework.

4.1 Traditional Essay Scoring Baselines

BERT Classifier. Similar to the model used in Han et al. [9], Yang et al. [44]'s work, we implemented a simple yet effective baseline model for score prediction based on BERT. This model integrated a fully connected prediction layer following the BERT output, and the BERT parameters remained unfrozen during training. Both the BERT model and the prediction layer were jointly trained on the training essay set (details in Appendix A.1).

SOTA baselines. We also incorporate models such as R^2BERT [44] and *Tran-BERT-MS-ML-R* [39], which represent the highest levels of performance in the ASAP AES task. These models serve as the high-level benchmarks against which we evaluate the performance of our LLM-based models.

4.2 Prompting LLMs

We considered various prompting strategies including with or without detailed rubrics context, zero-shot or few-shot settings. An illustrative example of a prompt and its corresponding model-generated output can be found in Table 9 in the Appendices.

GPT-4, zero-shot, without rubrics. In this setting, we simply provide the prompt and the target essay to GPT-4. The model then evaluates the essay and assigns a score based on its comprehension within the specified score range.

GPT-4, zero-shot, with rubrics. Alongside the prompt and the target essay, we also provide GPT-4 with explicit scoring rubrics, guiding its evaluation.

GPT-4, few-shot, with rubrics. In addition to the zero-shot settings, the few-shot prompts include sample essays and their corresponding scores. This assists GPT-4 in understanding the latent scoring patterns. With the given prompt, target essay, scoring rubrics, and a set of k essay examples, GPT-4 provides an appropriate score reflecting this enriched context.

- As indicated by prior studies in AES tasks [43], increasing the value of k did not consistently yield better results, showing a trend of diminishing marginal returns. Therefore, we choose a suitable $k = 3$ as described in the study.
- We explored two sampling approaches. The first involved randomly selecting essays from various levels of quality to help LLM understand the approximate level of the target essay. The second method adopted a retrieval-based approach, which has been proven to be effective in enhancing LLM performance [14, 27, 33]. Leveraging OpenAI's *text-embedding-ada-002* model, we calculated the embedding for each essay. This allowed us to identify the top k similar essays based on cosine similarity (excluding the target essay). Our experiments demonstrated that this retrieval strategy consistently yielded superior results. Therefore, we focused on the outcomes of this approach in the subsequent sections.

In all these configurations, we adopted the Chain-of-Thought (CoT) [42] strategy. This approach instructed the LLM to analyze and explain the provided materials before making final score determinations. Studies [15, 16, 46] have shown that this structured approach significantly enhances the capabilities of the LLM, optimizing performance in tasks that require inference and reasoning.

4.3 Fine-tuning LLMs

We conducted additional investigations into the effectiveness of supervised fine-tuning methods. Given that the ASAP and our CSEE datasets only include scores without expert explanations, we augmented these original datasets with explanations generated by GPT-4. To guide the explanation generation process, we provided GPT-4 with a few expert-curated explanations and a structured template. By organizing the data into an instructional format, we created fine-tuning inputs that enable the LLMs to not only generate accurate scores but also provide high-quality feedback.

---

[3] https://www.kaggle.com/c/asap-aes.

[4] Dataset is available in https://github.com/Xiaochr/LLM-AES

Table 1: The descriptive statistics of the ASAP dataset.

Essay Set	Essay Type	Grade Level	# of Essays	Avg. Length	Score Range
1	Persuasive/Narrative/Expository	8	1783	350	[2, 12]
2	Persuasive/Narrative/Expository	10	1800	350	[1, 6]
3	Source Dependent Responses	10	1726	150	[0, 3]
4	Source Dependent Responses	10	1772	150	[0, 3]
5	Source Dependent Responses	8	1805	150	[0, 4]
6	Source Dependent Responses	10	1800	150	[0, 4]
7	Persuasive/Narrative/Expository	7	1569	300	[0, 30]
8	Persuasive/Narrative/Expository	10	723	650	[0, 60]

Table 2: Descriptive statistics of our Chinese Student English Essay (CSEE) dataset.

Chinese Student English Essay Dataset	
# of schools	29
# of essay prompts	2
# of student essays	13,372
avg. essay length	124.74
avg. Overall score	10.72
avg. Content score	4.13
avg. Language score	4.05
avg. Structure score	2.55

We first fine-tuned OpenAI's *GPT-3.5-turbo*, one of the best-performing LLMs. However, due to the proprietary nature of GPT-3.5 and considerations such as data privacy, training and inference costs, and flexibility in fine-tuning, we also fine-tune an LLaMA3-8B [5] model with both original and augmented datasets. This recent open-source model mitigates these concerns and has remarkable capabilities, making it a more practical choice for use in educational scenarios. The fine-tuning details are presented in Appendix A.2.

4.4 Our Proposed Method

As previously mentioned, score prediction and explanation generation are distinct but interrelated tasks within the context of AES. Explanation generation, which covers the evaluation of content, language, and structure, necessitates deliberate and meticulous reasoning. On the other hand, score prediction can either be a swift process based on intuition and experiences or concluded after step-by-step inference. These features align with the idea of dual-process theory. Consequently, we have designed an integrated system that includes separate modules for each task: the Fast Module for quick score prediction and the Slow Module for detailed explanation generation. The framework of our proposed AES system is shown in Figure 1.

Slow Module: Fine-tuned LLM. The Slow Module forms the core of our AES system, capable of analyzing essays in depth, providing evidence based on specific rubrics, and deriving appropriate scores. This detailed process is time-intensive but yields valuable natural language reasoning that informs the final grading decision. In this study, we implemented the fine-tuned LLaMA3-8B as the Slow Module. It is worth noting that this module is interchangeable with any other qualified LLM, demonstrating the flexibility of our framework.

Fast Module: Lightweight Classifier. In many cases, swift score prediction is preferable to detailed reasoning. To reduce the time and computational cost associated with generating detailed explanations, we introduced a simple fully connected layer as a bypass before the initiation of text generation by the Slow Module. By using only the embeddings of the input text, the Fast Module not only conserves resources but also leverages the latent knowledge acquired during the fine-tuning of the Slow Module, aligning with the 'intuitive' facet of Fast Thinking.

When to switch from the Fast to Slow Thinking module is one of the challenges in the design of dual-process LLM. Previous frameworks employed heuristic rules or error feedback as the triggering criteria [17, 37], which might be impractical in real-world cases. Our Fast module also calculates the probabilities of each possible output score, which we standardize and treat as confidence scores. Predictions with low confidence are considered unreliable, triggering the Slow Module for self-reflection, or passing to external judges (either human or AI). This design aims to enhance essay scoring accuracy and foster effective human-AI collaboration, potentially elevating the complementary team performance.

For training, we first fine-tune the Slow Module using our explanation augmented dataset. Subsequently, we employ the Slow Module to derive input embeddings, which, paired with the rated scores, are used to train the Fast Classifier from scratch. During inference, essay inputs initially pass through the fine-tuned LLM and are transformed into the embedding format. They are then processed by the Fast Module to quickly derive scores. The Slow Module is activated only when prediction confidence is low or based on specific additional requirements.

5 Experimental Results

5.1 Performance of LLM-based Methods

We conducted experiments across all eight subsets of the ASAP dataset using both the LLM-based methods and baseline approaches. We adopted Cohen's Quadratic Weighted Kappa (QWK) as our primary evaluation metric, which is the most widely recognized automatic metric in AES tasks [28]. A higher QWK value indicates a greater degree of agreement between the predicted score and the

---

[5] https://llama.meta.com/llama3/

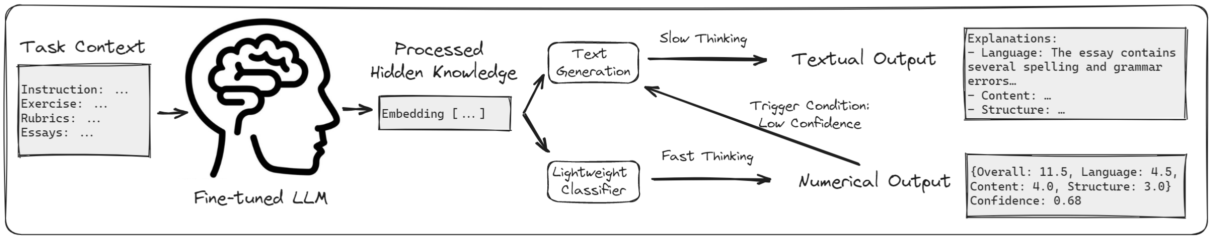

Figure 1: Our proposed Fast and Slow Thinking AES framework.

ground truth. For methods requiring a training dataset, we divided the data for each subset using an 80:20 split ratio between training and testing.

Our extensive experiments, as detailed in Table 3, revealed that despite using carefully curated prompts and providing detailed context, the zero-shot and few-shot capabilities of GPT-4 did not yield high QWK scores on the ASAP dataset. In zero-shot scenarios, GPT-4's performance was notably low, with some subsets scoring nearly as poorly as random guessing. For instance, Set 1 recorded a QWK of 0.0423 and Set 7 a QWK of 0.0809. This underperformance may be due to the broad scoring ranges and complex rubrics in ASAP, suggesting that even advanced LLMs like GPT-4 may struggle to fully comprehend and adhere to complicated human instructions. In few-shot settings, although there was an improvement in scoring performance, particularly for Sets 4-6, GPT-4 still significantly lagged behind SOTA grading methods. This is consistent with findings from recent studies that utilize LLMs for essay scoring.

When fine-tuned with the training dataset, the LLMs demonstrated significantly improved performance compared to the zero-shot and few-shot results, with QWK scores generally exceeding 0.7. However, these fine-tuned LLMs still did not surpass traditional SOTA methods. Within our framework, the performance of the fine-tuned open-source LLaMA3-8B was comparable to that of fine-tuned proprietary models. Even simple supervised fine-tuning (SFT) of LLaMA3 achieved notable results, suggesting that open-source LLMs might be a cost-effective choice for AES tasks. The findings from our CSEE dataset (see Table 4) align with those on the ASAP dataset, indicating that our framework predicts reliable scores across content, language, and structure dimensions.

Although LLMs do not match traditional methods in terms of scoring accuracy, they excel at generating detailed explanations, benefiting both educators and students. Notably, when trained to produce both scores and explanations in a single output (our proposed Slow Module), LLaMA3-8B experienced a performance drop in grading accuracy. This decrease may be attributed to the model's optimization process, where numerical score values are treated similarly to textual data in the output, leading to suboptimal accuracy. In our Fast and Slow Thinking framework, however, separating numerical from textual outputs and integrating them based on a trigger condition improved the QWK scores, enhancing collaborative performance. Additionally, we evaluated the quality of explanations generated by our AES system against those produced by GPT-4. Through a comparison competition among crowdsourced workers, analyzing 20 sets of paired essay grading explanations, our system achieved a win rate of 35%, a tie rate of 40%, and a loss rate of 25%. These results demonstrate that our explanations are of high quality and comparable to those generated by GPT-4.

5.2 Further Analyses

Consistency. To assess the consistency of scores predicted by LLM-based methods, we conducted the same experiment five times, each with the *temperature* parameter of the LLMs set to 0. We observed that over 80% of the ratings remained unchanged across these trials, indicating a high level of consistency. We then computed the average of these three values to determine the final results.

Generalizability. The eight subsets of the ASAP dataset, featuring diverse scoring criteria and ranges, serve as an excellent framework for evaluating the generalization capabilities of models. For methods such as fine-tuning and traditional baselines that require training data, we first trained the models on one subset and then assessed their performance across the remaining datasets. For example, we trained on Set 1 and tested on Sets 2-8, keeping the model weights fixed. We selected fine-tuned GPT-3.5 and the BERT Classifier to represent LLM-based and traditional methods, respectively. As detailed in Table 5, our fine-tuned GPT-3.5 generally outperformed the BERT classifier, although there were instances of underperformance, notably when trained on Set 4 and tested on Sets 1 and 7. The BERT classifier showed particularly weak generalization when trained on Sets 7 and 8, performing close to random guessing.

Prediction Confidence and Error Analysis. To assess the reliability of confidence scores, we segmented the test samples based on the output confidence and observed a strong correlation between these scores and model performance in Figure 2. The trigger condition for switching from the Fast to the Slow Module is set when the confidence score falls below 0.2. Although the Slow Module generally exhibits lower performance compared to the Fast Module, the overall performance of the integrated system improved. We further investigated the error cases of the Fast Module to see how the Slow Thinking contributes to the complementary performance. A typical example is: For essays whose contents deviate from the prompt, the Fast Module tends to predict a higher score due to the full structure and appropriate language, while the LLM-based Slow Module identifies these discrepancies, providing scores closer to the ground truth with natural language explanations for the irrelevant contents. Thus, the error analysis also supports the idea that employing detailed reasoning for cases with low confidence is an effective grading strategy.

Table 3: Comparison of QWK scores for LLM-based methods and the baselines under the ASAP dataset. The "Ex." column indicates whether the model output includes natural language explanations alongside the predicted scores.

	Ex.	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Avg.
BERT Classifier	✗	0.6486	0.6284	0.7327	0.7669	0.7432	0.6810	0.7165	0.4624	0.6725
Tran-BERT-MS-ML-R	✗	0.8340	0.7160	0.7140	0.8120	0.8130	0.8360	0.8390	0.7660	0.7910
R^{2}BERT	✗	0.8170	0.7190	0.6980	0.8450	0.8410	0.8470	0.8390	0.7440	0.7940
GPT-4, zero-shot, w/o rubrics	✓	0.0423	0.4017	0.2805	0.5571	0.3659	0.5021	0.0809	0.4188	0.3312
GPT-4, zero-shot, with rubrics	✓	0.0715	0.3003	0.3661	0.6266	0.5227	0.3448	0.1101	0.4072	0.3437
GPT-4, few-shot, with rubrics	✓	0.2801	0.3376	0.3308	0.7839	0.6226	0.7284	0.2570	0.4541	0.4743
Fine-tuned GPT-3.5	✗	0.7406	0.6183	0.7041	**0.8593**	0.7959	**0.8480**	**0.7271**	**0.6135**	**0.7384**
Fine-tuned LLaMA3	✗	0.7137	**0.6696**	0.6558	0.7712	0.7452	0.7489	0.6938	0.2952	0.6617
Ours (based on LLaMA3)	✓	**0.7612**	0.6517	**0.7238**	0.8093	**0.8118**	0.7764	0.7071	0.4885	0.7162
Fast module	✗	0.7580	0.6395	0.7228	0.7995	0.8023	0.7753	0.7157	0.5075	0.7151
Slow module	✓	0.6048	0.5621	0.5700	0.6992	0.6774	0.5943	0.5772	0.2677	0.5691

Table 4: Comparison of QWK scores for LLM-based methods and the baselines under our CSEE dataset. The "Ex." column indicates whether the model output includes natural language explanations alongside the predicted scores.

	Ex.	Overall	Content	Language	Structure
BERT Classifier	✗	**0.7674**	0.7312	0.7203	0.6650
GPT-4, zero-shot, w/o rubrics	✓	0.4688	0.4412	0.3081	0.5757
GPT-4, zero-shot, with rubrics	✓	0.5344	0.5391	0.4660	0.4256
GPT-4, few-shot, with rubrics	✓	0.6729	0.6484	0.6278	0.4661
Fine-tuned GPT-3.5	✗	0.7532	0.7241	**0.7513**	0.6576
Fine-tuned LLaMA3	✗	0.7544	0.7321	0.7084	0.6461
Ours	✓	0.7634	**0.7347**	0.7192	**0.6656**
Fast module	✗	0.7364	0.7272	0.7072	0.6627
Slow module	✓	0.7310	0.6810	0.6990	0.6412

Time Efficiency. Training the Slow Module for each epoch with our explanation-augmented dataset requires around 2 hours using an RTX 4090 24GB GPU, and the inference process consumes about 10 GPU hours. Meanwhile, training the Fast Module takes less than 0.5 hours, and inferences are completed in just 0.2 hours. Our proposed framework, which incorporates a confidence trigger condition, offers an effective trade-off by enhancing both accuracy and efficiency.

6 Human-AI Co-Grading Experiment

Given that the AES system not only provides score predictions but also functions as a teammate to educators, we further explore the effectiveness of our proposed system in assisting human grading.

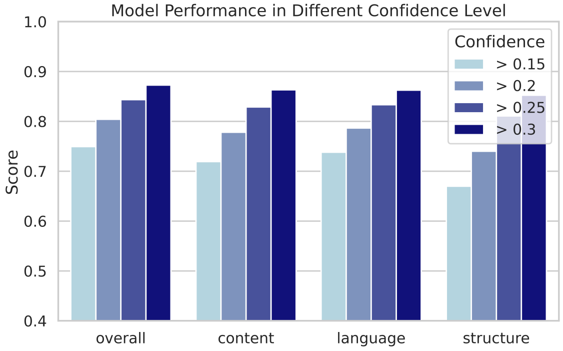

Figure 2: QWK scores of our Fast module in different confidence levels.

6.1 Experiment Design

To investigate the performance of human-only, AI-only, and human-AI team collaboration, we conducted a two-stage within-group experiment. We randomly selected 50 essays from the test set of our CSEE dataset, all on the same topic. We recruited 10 college students from a Normal University in Beijing, who are prospective high school teachers with no current grading experience, to serve as novice evaluators. The novices had a male-to-female ratio of 4:6, with ages ranging from 19 to 23 years (from freshmen to seniors). Their English capabilities were certified by China's College English Test (CET). None of the novices have the experience of grading student essays currently. Additionally, 5 experienced high school English teachers participated as expert evaluators, with teaching tenures ranging from 8 to 20 years.

Before evaluation, all participants received training on the standard scoring rubrics. They were also incentivized with appropriate remuneration for their participation. The instructions for the evaluators include the standard scoring rubrics of the College Entrance

Table 5: Generalization comparison of QWK scores for the Fine-tuned GPT-3.5 and the BERT Classifier under the ASAP dataset.

		Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
Trained on Set 1	BERT Classifier	-	0.3299	0.1680	0.1380	0.3045	0.1234	0.3002	0.1541
	Fine-tuned GPT-3.5	-	**0.5216**	**0.5405**	**0.4891**	**0.5076**	**0.6344**	**0.6306**	**0.3126**
Trained on Set 2	BERT Classifier	0.2776	-	0.1975	0.2392	0.1750	0.1453	0.2474	0.3783
	Fine-tuned GPT-3.5	**0.4270**	-	**0.4131**	**0.4619**	**0.5958**	**0.5579**	**0.5438**	**0.6684**
Trained on Set 3	BERT Classifier	0.3468	**0.4444**	-	0.6230	0.6319	0.5299	0.4368	**0.2427**
	Fine-tuned GPT-3.5	**0.3991**	0.2488	-	**0.7674**	**0.7714**	**0.7150**	**0.4964**	0.1134
Trained on Set 4	BERT Classifier	**0.3257**	**0.5332**	**0.6267**	-	0.5483	0.4959	**0.4659**	0.3204
	Fine-tuned GPT-3.5	0.0631	0.3493	0.4908	-	**0.6515**	**0.7420**	0.0865	**0.3419**
Trained on Set 5	BERT Classifier	0.4051	0.3341	0.4264	0.4202	-	0.5243	**0.3255**	0.2035
	Fine-tuned GPT-3.5	**0.4354**	**0.4301**	**0.5765**	**0.6877**	-	**0.7368**	0.1061	**0.3118**
Trained on Set 6	BERT Classifier	**0.3164**	0.3462	0.4000	0.3067	**0.4882**	-	**0.2303**	**0.3047**
	Fine-tuned GPT-3.5	0.1342	**0.3607**	**0.4579**	**0.3157**	0.3734	-	0.0061	0.0859
Trained on Set 7	BERT Classifier	0.0975	0.0086	0.1854	0.0328	0.0554	0.1244	-	**0.2917**
	Fine-tuned GPT-3.5	**0.5862**	**0.3993**	**0.4865**	**0.4425**	**0.4494**	**0.4417**	-	0.2157
Trained on Set 8	BERT Classifier	0.0560	0.1102	0.0110	0.0164	0.0371	0.0454	0.1777	-
	Fine-tuned GPT-3.5	**0.2714**	**0.4822**	**0.4768**	**0.6009**	**0.4199**	**0.3231**	**0.5460**	-

Examination in China and several grading examples. The writing exercise and the essays designated for assessment will be presented to the evaluators. Moreover, supplementary feedback (scores, output confidences, and explanations) will be provided for the experimental groups. To enhance the evaluators' comprehension and avoid possible misunderstandings, all the information provided has been translated into Chinese.

During the experiment, all evaluators first graded the essays independently using standard rubrics. Subsequently, they were provided with the scores, prediction confidence levels, and explanations generated by our AES system and had the option to revise their initial scores based on this augmented information. To gather feedback on the process, we distributed questionnaires where evaluators rated their experience on a 5-point Likert scale, with higher scores indicating better-perceived performance.

On the other hand, for an intriguing comparative analysis, we fed the scores and explanations provided by our Fast and Slow framework back to a general LLM (GPT-4) for reflection. From our previous computational experiments, we know that the grading performance of our model is superior to general-purpose LLMs. Here, our objective is to explore whether general LLMs can critically evaluate supplementary reference information and improve their grading performance in a manner similar to human evaluators. This investigation aligns with concepts such as self-reflection [11, 40], LLM-as-a-Judge [8, 45], and the mixture-of-experts [32] in the LLM literature.

In this study, we address the following research questions:

- Can novice and expert human evaluators achieve complementary performance in terms of accuracy and efficiency using the proposed AES system and collaborative workflow?
- Can general LLM evaluators leverage superior AES feedback generated by an advanced LLM to achieve complementary grading performance?
- Does the incorporation of prediction confidence and explanation generation contribute to performance improvements?

6.2 Results

Feedback generated by LLM elevates novice evaluators to expert level. As depicted in Figure 3 and Table 6, our findings reveal that novice graders, with the assistance of LLM-generated feedback (including both scores and explanations), achieved a significant improvement in performance. Their average QWK improved from 0.5256 to 0.6609, with a p-value of less than 0.01. Furthermore, when comparing the performance of LLM-assisted novice evaluators (mean QWK of 0.6609) to that of expert graders (mean QWK of 0.7117), no statistical difference was found between the two groups (p-value = 0.27). This indicates that with LLM support, novice evaluators achieved a level of grading proficiency comparable to that of experienced experts. Similar trends were observed in the scores for the Content and Language dimensions. The expert evaluators noted that the Structure dimension is the most ambiguous and difficult part of the grading task, resulting in the lowest QWK values among the three dimensions for both human and LLM evaluators.

LLM-generated feedback enhances expert efficiency and consistency. Integrating LLM-generated feedback into the expert grading process increased the average Quadratic Weighted Kappa (QWK) from 0.7117 to 0.7734, surpassing the performance of AES systems (QWK of 0.7302) on these essay samples. This demonstrates superior complementary performance, where the Human-AI team outperforms both individual human graders and the AI. Although this improvement is not statistically significant (p-value = 0.13), the

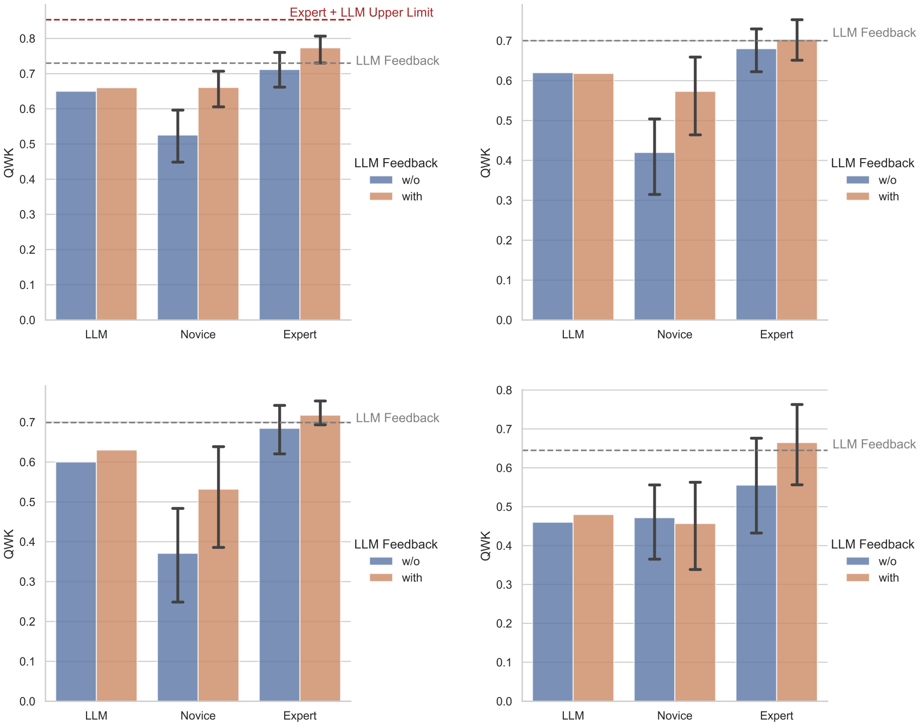

Figure 3: LLM-assisted grading experiment results for the novice, expert, and GPT-4 evaluators. From the top left to the bottom right is the result of the Overall score, Content score, Language score, and Structure score, respectively.

Table 6: *t*-test of QWK scores for different experimental groups. *Diff.* means the difference of means between the two groups of QWK scores.

	Diff.	*t* statistic	*p*-value
Expert vs. Novice	**0.1860*****	3.2152	**0.0068**
Novice+LLM vs. Novice	**0.1353*****	2.8882	**0.0098**
Expert+LLM vs. Expert	0.0617	1.7128	0.1251
Novice+LLM vs. Expert	-0.0508	-1.1566	0.2682

benefits of LLM augmentation for experts were evident in other key areas.

In terms of grading efficiency, we tracked the time required for independent grading and the subsequent revision process with LLM feedback. Due to the within-subjects design, independent grading was naturally more time-consuming, as participants had to read and evaluate the essays entirely on their own, whereas the revision process involved primarily reviewing the LLM-predicted scores. This design limits direct comparisons between expert-only and expert + LLM grading times. To address this, we relied on self-reported questionnaires. As summarized in Table 7, experts perceived the grading process to be less time-consuming when assisted by the LLM.

Table 7: Experts' feedback after grading student essays with the support of the LLM-based system.

Metrics	**Score**
Perceived accuracy of LLM overall score	**4.3**/5
Perceived accuracy of LLM content score	**4.0**/5
Perceived accuracy of LLM language score	3.9/5
Perceived accuracy of LLM structure score	3.8/5
Helpfulness of the predicted scores	**4.6**/5
Helpfulness of the confidence scores	**4.8**/5
Helpfulness of LLM explanations	**4.7**/5
Efficiency of LLM assistance	**4.4**/5
Willingness to use our AES system	**4.3**/5

We also compared the completion times of experts and novices under LLM assistance, finding that expert + LLM grading was significantly faster than novice + LLM grading (39 seconds vs. 63 seconds per exercise). Moreover, the data revealed that experts spent less time on questions where their scores aligned closely with the LLM's predictions, but spent more time deliberating on cases where discrepancies arose. Unlike novices, experts were less likely to alter their initial scores in such instances, reflecting confidence in their expertise. However, in some cases, this reluctance to revise scores could limit the potential improvements offered by accurate LLM predictions.

Furthermore, a reduction in the standard deviation of expert ratings was observed, indicating a higher level of consensus among experts. This suggests that LLM-assisted grading leads to more consistent evaluations of student essays. Experienced domain experts also commended the accuracy and practicality of the LLM-generated feedback in the survey and interviews, particularly praising the prediction confidence mechanism which alerted them to scrutinize more challenging cases. These findings demonstrate the potential of augmenting human grading processes with LLMs, paving the way for more efficient, consistent, and scalable evaluations for real-world learning analytics.

Human evaluators outperform LLM evaluators when it comes to effectively leveraging high-quality feedback. Despite a carefully crafted prompt for the general LLM evaluator to utilize additional feedback, there was no significant improvement in its performance (see Figure 3). On the contrary, as previously discussed, both novice and expert human evaluators effectively utilized superior LLM-generated feedback to enhance their grading in terms of accuracy, consistency, and efficiency. This indicates that while GPT-4 can match expert-level grading in general performance, its capacity to critically improve with additional information is limited compared to human evaluators' reflective and adaptive abilities. This observation highlights the potential need for more advanced approaches that enable LLMs to better integrate additional feedback information for improved performance.

The importance of prediction confidence and explanations. We previously assessed the reliability of prediction confidence from our Fast Module and noted a modest improvement in model performance after self-reflection by the Slow Module (as shown in the LLM column of Figure 3). In the context of human-AI collaboration, we focused on cases where the predicted scores presented to human evaluators were of low confidence (below 0.2). We observed that the overall QWK scores for expert and novice evaluators were 0.6809 and 0.5680. These QWK values, lower than the average human performances, suggest that these essays are inherently challenging to grade, even for humans. However, human performances exceeded that of the LLM Slow Module's 0.5478 QWK, achieving complementary team performance. These findings support a practical, intuitive LLM-assisted decision-making workflow: the model manages routine cases with high confidence and minimal human intervention, while low-confidence cases are presented to human collaborators for in-depth analysis and final decision-making.

Challenges in fully utilizing additional information by human and LLM evaluators. In our analysis, we established the upper bound for the utilization of additional information as the scenario where the most advantageous outcomes are consistently chosen in the combined ratings. For example, an expert evaluator using LLM-generated feedback reaches her performance upper bound (a QWK of 0.8535) by adopting superior LLM advice where applicable, and relying on her own judgments when they are closer to the ground truth. Our findings indicate that none of the three experiment groups (human novices with LLM feedback, human experts with LLM feedback, and general LLM with superior feedback) managed to surpass the performance of their respective upper bounds. Notably, novices altered their initial scores in 69.5% of cases, yet only 60.0% of these modifications resulted in improved performance, with the rest being counterproductive. Conversely, the experts, displaying greater confidence in their judgments, adhered to LLM-generated feedback in just 26.0% of instances, but 84.6% of these adjustments were beneficial, bringing them closer to the ground truth scores. This raises compelling questions about why both humans and LLMs hardly outperform the superior reference scores and how more nuanced design strategies might elevate their performance, presenting a fascinating direction for future research.

7 Conclusion

In this study, we explored the capabilities of LLMs within AES systems. With detailed contexts, clear rubrics, and high-quality examples, GPT-4 demonstrated satisfactory performance, consistency, and generalizability. Further accuracy enhancements were achieved through supervised fine-tuning using task-specific instruction datasets, bringing LLM performance close to conventional SOTA methods. To leverage the LLMs' ability to generate natural language explanations along with predicted scores, we introduced an open-source Fast and Slow Thinking AES framework. This framework not only matches the quality of proprietary models but also offers greater efficiency.

Our research extended into human-AI co-grading experiments within this new framework. A notable finding was that LLMs not only automated the grading process but also augmented the grading skills of human evaluators. Novice graders, with support from our AES framework, reached accuracy levels comparable to those of experienced graders, while expert graders showed gains in efficiency and consistency. The collaboration between humans and AI particularly enhanced performance in handling low-confidence cases, demonstrating a significant synergy that approached the upper limits of team performance. These results highlight the transformative potential of AI-assisted and human-centered decision-making workflows, especially in elevating those with limited domain knowledge to expert-level proficiency. This study illuminates promising future directions for human-AI collaboration and underscores the evolving role of AI in educational contexts.

This study still has certain limitations. Firstly, although our CSEE dataset includes a substantial number of student essays, these essays originate from only two final exams designed for high school English learners in China. This raises concerns about the robustness of our proposed AES system when applied to a broader range of topics and diverse student demographics. Secondly, our human-AI collaboration experiment, while indicative of promising directions for future human-AI co-grading tasks, is a pilot study that yields general results. Further experiments are necessary to thoroughly explore the mechanisms of complementary team performance, such as identifying circumstances under which humans are likely to recognize and correct their errors following AI feedback, or instances where unreliable AI feedback could potentially mislead them. A deeper understanding of these collaboration mechanisms will enable researchers to develop AES systems that offer more effective support to educators.

Ethical Considerations

We secured Institutional Review Board (IRB) approval for both the data collection and the human-AI co-grading experiment (details are provided in our GitHub repository). In our CSEE dataset, all personal information concerning the students has been anonymized to safeguard their privacy. The dataset comprises solely of essays and the corresponding scores, omitting any additional information that might raise ethical concerns. However, details of the data annotation process remain undisclosed to us, including the number of teachers involved in the scoring and the level of inter-annotator agreement among them. We have also obtained explicit consent to use the data exclusively for research purposes from both teachers and students.

References

[1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[2] Samuel C Bellini-Leite. 2023. Dual Process Theory for Large Language Models: An overview of using Psychology to address hallucination and reliability issues. *Adaptive Behavior* (2023). https://doi.org/10.1177/10597123231206604

[3] Martin Benfeghoul, Umais Zahid, Qinghai Guo, and Zafeirios Fountas. 2024. When in Doubt, Think Slow: Iterative Reasoning with Latent Imagination. *arXiv preprint arXiv:2402.15283* (2024).

[4] Eujene Nikka V. Boquio and Prospero C. Naval, Jr. 2024. Beyond Canonical Fine-tuning: Leveraging Hybrid Multi-Layer Pooled Representations of BERT for Automated Essay Scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 2285–2295. https://aclanthology.org/2024.lrec-main.204

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 1072–1077.

[7] Eva Eigner and Thorsten Händler. 2024. Determinants of LLM-assisted Decision-Making. *arXiv preprint arXiv:2402.17385* (2024).

[8] Fabricio Goes, Zisen Zhou, Piotr Sawicki, Marek Grzes, and Daniel G Brown. 2022. Crowd score: A method for the evaluation of jokes using large language model AI voters as judges. *arXiv preprint arXiv:2212.11214* (2022).

[9] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. 2023. FABRIC: Automated Scoring and Feedback Generation for Essays. *arXiv preprint arXiv:2310.05191* (2023).

[10] Kenneth Holstein and Vincent Aleven. 2022. Designing for human–AI complementarity in K-12 education. *AI Magazine* 43, 2 (2022), 239–248.

[11] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1827–1843. https://doi.org/10.18653/v1/2023.findings-emnlp.123

[12] D. Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux. https://books.google.com.tr/books?id=AV9x8XakdV0C

[13] Mahammed Kamruzzaman and Gene Louis Kim. 2024. Prompting Techniques for Reducing Social Bias in LLMs through System 1 and System 2 Cognitive Processes. *arXiv preprint arXiv:2404.17218* (2024).

[14] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HklBjCEKvH

[15] Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context?. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 537–563. https://doi.org/10.18653/v1/2022.findings-emnlp.38

[16] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making Language Models Better Reasoners with Step-Aware Verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5315–5333. https://doi.org/10.18653/v1/2023.acl-long.291

[17] Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2024. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems* 36 (2024).

[18] Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. 2020. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13389–13396.

[19] Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

[20] Sandeep Mathias and Pushpak Bhattacharyya. 2018. Thank "goodness"! a way to measure style in student essays. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*. 35–41.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[22] Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering* 10, 1 (2004), 25–55.

[23] Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics* 2, 2 (2023), 100050.

[24] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 394–403. https://doi.org/10.18653/v1/2023.bea-1.32

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[27] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083* (2023).

[28] Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review* 55, 3 (2022), 2495–2527.

[29] Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring. *arXiv preprint arXiv:2008.01441* (2020).

[30] Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482* (2019).

[31] Yafet Salim, Valdi Stevanus, Edwardo Barlian, Azani Cempaka Sari, and Derwin Suhartono. 2019. Automated English digital essay grader using machine learning. In *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*. IEEE, 1–6.

[32] Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, et al. 2023. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv preprint arXiv:2305.14705* (2023).

[33] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652* (2023).

[34] Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. *arXiv preprint arXiv:2404.15845* (2024).

[35] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1070–1075.

[36] Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 1882–1891.

[37] Xiaoyu Tian, Liangyu Chen, Na Liu, Yaxuan Liu, Wei Zou, Kaijiang Chen, and Ming Cui. 2023. DUMA: a Dual-Mind Conversational Agent with Fast and Slow Thinking. *arXiv preprint arXiv:2310.18075* (2023).

[38] Karel van den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. 2019. Six challenges for human-AI Co-learning. In *Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*. Springer, 572–589.

[39] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3416–3425. https://doi.org/10.18653/v1/2022.naacl-main.249

[40] Ziqi Wang, Le Hou, Tianjian Lu, Yuexin Wu, Yunxuan Li, Hongkun Yu, and Heng Ji. 2023. Enable Language Models to Implicitly Learn Self-Improvement From Data. *arXiv preprint arXiv:2310.00898* (2023).

[41] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.

[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[43] Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating Short L2 Essays on the CEFR Scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 576–584. https://doi.org/10.18653/v1/2023.bea-1.49

[44] Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1560–1569.

[45] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).

[46] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=WZH7099tgfM

A Model Training Details

A.1 Fine-tuning BERT Classifier Baseline

We employed the *bert-base-uncased* BERT model from the huggingface transformers library[6] using PyTorch. A simple fully connected layer was added to perform the classification task. The datasets were divided into training and testing sets at an 8:2 ratio. To ensure better reproducibility, we set all random seeds, including those for dataset splitting and model training, to the value 42. During training, we used cross-entropy loss as our loss function. We allowed BERT parameters to be fine-tuned, without freezing them, in line with the objective function. AdamW was chosen as the optimizer, with a learning rate set to 10^{-5} and epsilon at 10^{-6}. With a batch size of 16 and a maximum of 10 training epochs, we also integrated an early stopping strategy to mitigate potential overfitting. All the experiments of the BERT baseline were run with 2 RTX A4000 16G GPUs in around one week.

A.2 Fine-tuning LLaMA3

We fine-tuned the *llama-3-8b-bnb-4bit* model using the *unsloth* framework[7]. For this process, we employed a Parameter-Efficient Fine-Tuning (PEFT) approach with a rank of 16 and a LoRA alpha value of 16. We utilized an 8-bit AdamW optimizer, starting with an initial learning rate of 2×10^{-4}. After 50 warm-up steps, the learning rate was scheduled to decay linearly, with the weight decay parameter set at 0.01. We maintained all random seeds at 3407 and completed the fine-tuning over 2 epochs. All experiments involving the fine-tuned LLaMA3-8B were conducted using a single RTX 4090 24GB GPU, spanning approximately three weeks.

B LLM Prompts

The prompts used for LLMs in our study fall into two distinct categories: firstly, the zero-shot and few-shot configurations of GPT-4; secondly, the instructions for fine-tuning and inference of GPT-3.5 and LLaMA3-8B. The prompts for the few-shot scenario incorporate those used in the zero-shot setting and overlap with the fine-tuning prompts. Therefore, for clarity and conciseness, we present examples of the *GPT-4, few-shot, with rubrics* and the inputs of fine-tuned LLaMA3-8B in Table 9.

---

[6] https://huggingface.co/docs/transformers/

[7] https://github.com/unslothai/unsloth

Table 8: Rubrics for evaluating high school student essays in our private dataset.

Rubrics

Overall Score (20 points) = **Content Score** (8 points) + **Language Score** (8 points) + **Structure Score** (4 points)

Content Dimension (8 points in total)

- 6-8 points:
    - Content is complete with appropriate details
    - Expression is closely related to the topic
- 3-5 points:
    - Content is mostly complete
    - Expression is fundamentally related to the topic
- 0-2 points:
    - Content is incomplete
    - Expression is barely related or completely unrelated to the topic

Language Dimension (8 points in total)

- 6-8 points:
    - Language is accurate with diverse sentence structures and little or no errors (2 errors or fewer, 8 points; 3-4 errors, 7 points; 5-6 errors, 6 points)
    - Language expression is mostly appropriate
- 3-5 points:
    - Language is not quite accurate, with some variation in sentence structures and several errors, but they don't impede understanding (7-8 errors, 5 points; 9-10 errors, 4 points; 11-12 errors, 3 points)
    - Language expression is somewhat inappropriate
- 0-2 points:
    - Language is hopelessly inaccurate with numerous language errors, hindering understanding (more than 12 errors)
    - Language expression is completely inappropriate

Structure Dimension (4 points in total)

- 3-4 points:
    - Clearly and logically structured
    - Smooth and coherent transitions
- 1-2 points:
    - Mostly clearly and logically structured
    - Relatively smooth and coherent transitions
- 0-1 points:
    - Not clearly and logically structured
    - Fragmented and disconnected structures and sentences

Table 9: The example prompt (the *GPT-4, few-shot, with rubrics* approach and the fine-tuned LLaMA3-8B) and the corresponding model output.

Example Prompt	Example Output
As a virtual evaluator with expertise in English composition, your role is to critically analyze and grade student essays according to a predetermined set of rubrics. You are to act as an impartial judge and evaluate the essays based on the quality of the writing and adherence to the essay prompt. Here are the specific guidelines for each score: {*Rubrics*} Sample Essay Prompt: {*Essay Prompt*} The graded example essays: {*Examples*} Student's Essay to Evaluate: {*Target Student Essay*} Task Breakdown: 1. Carefully read the provided essay prompt, scoring guidelines, and the student's essay. 2. In the Explanations part, identify specific elements in the essay referring to the rubrics. In the language dimension, list all the spelling and grammar errors, and count the number of them to determine the Language Score. The Explanations for each dimension should be as detailed as possible. 3. Determine the appropriate scores according to the analysis above. Please present your evaluation in the following manner: Explanations: ..., Content Score: ... Explanations: ..., Language Score: ... Explanations: ..., Structure Score: ... Explanations: ..., Total Score: ... Your final evaluation: [Total Score: ..., Content Score: ..., Language Score: ..., Structure Score: ...]	Explanations: The student's essay responds accurately to Jim's request for suggestions on how to learn Chinese, providing thoughtful advice and support. Detailed suggestions include listening to Chinese radio, reading Chinese literature, and using mobile applications. The tone is hopeful and supportive. Content Score: 8 Explanations: There are no language errors; the grammar and spelling are all accurate. The student used formal, respectful language suitable for this kind of communication and also included diversity in sentence structures. Language Score: 8 Explanations: The essay is well-structured and organized. It follows the accepted structure of a reply email, with an opening greeting, a body containing the main content, and a closing remark. Transitions are also smooth and coherent, logically linking one thought to another. Structure Score: 4 Total Score: This is the sum of the content, language, and structure scores. Total Score: 20 Your final evaluation: [Total Score: 20, Content Score: 8, Language Score: 8, Structure Score: 4]
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: You are an experienced English teacher analyzing high school students' essays according to a specific rubric. Evaluate the following essay based on three dimensions: Content, Language, and Structure, and provide the overall assessment. Please provide your evaluation in the following JSON format: {*Output Format*} ### Input: Scoring rubrics: {*Rubrics*} Essay Prompt: {*Essay Prompt*} Student's Essay to Evaluate: {*Target Student Essay*} ### Response: {}	{ content: {completeness: The essay covers the basic requirements of the prompt by introducing design ideas for the graduation ceremony and explaining the reasons behind these designs. However, the explanations are quite brief and lack depth., topic_relevance: The essay is related to the given topic, as it discusses ideas for a graduation ceremony design., content_details: The details provided are minimal and lack depth. For example, the student mentions launching a ball in the dining hall and writing gratitude letters but does not elaborate on how these activities contribute to the overall design or the reasons behind these choices in detail., score_level: Level 2, score_point: 4.5}, language: {...}, structure: {...}, overall: {...} }