WrAFT: A Modular Large Language Model-Powered Automated Writing Evaluation System for Argumentative Essays

Anonymous ACL submission

Abstract

This study presents *WrAFT*, a **Wr**iting **A**ssessment and **F**eedback **T**ool, that delivers both accurate and reliable scores and effective comprehensive feedback to argumentative essays. *WrAFT* adopts a modular design by dividing the automated writing evaluation (AWE) tasks into scoring, surface-level feedback and deep-level feedback modules. In building the system, we evaluated various large language models (LLMs), including LLaMA-3.3-70B-Instruct, GPT-4o, and Claude 3.7, through both direct prompting and supervised fine-tuning approaches. A proprietary dataset of 480 TOEFL Independent Writing essays with official benchmark scores was utilized. Our evaluation demonstrates that *WrAFT* achieves state-of-the-art performance in scoring with a quadratic weighted kappa (QWK) of 0.84 and an root mean square error (RMSE) of 0.44 against benchmark scores on a score scale of 0-5. System-generated feedback also receives high approval ratings from human evaluators (96.14% for surface-level, 93.03% for deep-level macro feedback, and 94.69% for deep-level micro feedback). An interactive user interface has been developed for the system, publicly available and free to use.

1 Introduction

Argumentative writing is emphasized in secondary and post-secondary curricula as it fosters higher-order thinking (Graff, 2003; Kuhn, 2005). Assessing such open-ended writing reliably, however, is a notoriously difficult task for human raters: it demands considerable time, and even trained evaluators can exhibit subjective biases and inconsistency in their judgments (Shermis and Burstein, 2003). Automated writing evaluation (AWE) systems have thus emerged as a promising solution to score and provide feedback on student essays at scale.

Traditional AWE systems have mostly focused on scoring (Li and Ng, 2024), either through linguistic features, shallow text similarity measures (Li and Wu, 2023) or through neural network approaches to model input essays, giving grades based on a single vector representation of the essay (Dong et al., 2017; Jin et al., 2018). Though some of the systems demonstrated satisfactory scoring agreements with human raters, such as e-Rater[1], the scoring engine of Criterion, Education Testing Service's (ETS) AWE tool for GRE and TOEFL writing (Attali and Burstein, 2006), they often failed to capture higher-order thinking in writing such as coherence or argumentation.

The emergence of powerful large language models (LLMs) opens the door to AWE systems that not only predict a score but also give rich explanatory feedback on content. However, a comprehensive review of AWE research by Li and Ng (2024) shows that in recent years, this area of research seems to be narrowly focused on developing a sophisticated model that can beat competing models in scoring a standard evaluation dataset, such as the Automated Student Assessment Prize (ASAP[2]) and the Cambridge Learner Corpus-First Certificate in English exam (CLCFCE; Yannakoudakis et al., 2024), while there is a lack of feedback generation and validation. Li and Ng (2024) proposes that there should be different layers to AWE systems, including holistic or trait-specific scores, written feedback, and essays revised by experts. This proposal coincides with the need for comprehensive feedback stressed by second language writing research, which typically includes both corrective edits to surface-level errors in grammar, wording and mechanics, and deep-level feedback comments on issues such as organization, coherence and argumentation (Bitchener and Knoch, 2008). While some researchers caution against the potential cognitive overload that comprehensive feedback might impose on learners (Truscott, 1996), comprehensive

[1] https://www.ets.org/erater/about.html

[2] https://www.kaggle.com/datasets/lburleigh/asap-2-0

feedback remains a prevalent pedagogical practice in language education, as learners may selectively attend to aspects they find most relevant. However, existing AWE systems tend to provide either surface-level feedback, such as grammar error correction (GEC) systems (Imamura et al., 2012; Bryant et al., 2017; Rozovskaya and Roth, 2019; Grundkiewicz and Junczys-Dowmunt, 2019), or deep-level feedback in a generic manner (e.g., only pointing out that "the essay could use a clearer thesis and more examples") (Liu and Kunnan, 2015; Ranalli and Yamashita, 2021). In addition, feedback comments are frequently consolidated in standalone paragraphs separate from the essay text, as observed in (Stahl et al., 2024), which makes it inconvenient for users to associate the feedback with specific elements of their writing.

In this study, we built an LLM-based AWE system for argumentative essays that provides accurate and reliable scoring and comprehensive feedback through an interactive UI. Drawing on the first two layers proposed by Li and Ng (2024), we adopted a modular architecture with separate modules for scoring, surface-level feedback and deep-level feedback. For this research, we obtained from Education Testing Service (ETS) a proprietary dataset of 480 TOEFL Independent Writing essays with benchmark scores on the original scale of 0-5. The contribution of our research is as follows:

- We are the first to utilize a proprietary TOEFL writing dataset in building an AWE system that delivers both accurate and reliable scores aligned with ETS benchmarks and comprehensive feedback. The scoring module of the system achieves state-of-the-art (SOTA) performance with a QWK of **0.84** and an RMSE of **0.44** on a 0-5 score scale .

- We show that for feedback generation, a task with non-deterministic output, supervised fine-tuning was less effective than directly prompting SOTA LLMs to elicit feedback on specific traits/aspects in writing.

- Human validation of the comprehensive feedback generated by our system suggests highly satisfactory performance in correcting surface-level grammatical and mechanical errors, as well as in commenting on macro structure and micro aspects including context-dependent grammar, clarity, coherence, argumentation, and formality.

- We built an interactive user interface to visualize comprehensive feedback through in-line corrective edits and anchored comments.

2 Related work

2.1 Early automated writing evaluation

Automated writing evaluation (AWE) is defined as "the process of evaluating and scoring written prose via computer programs" (Shermis and Wilson, 2024, p.1), which includes for the dual tasks of scoring and feedback. AWE has a long history in NLP and education research, dating back to the seminal work of Page (1967) who first outlined the possibility of grading essays by computer and developed Project Essay Grade (Page, 2003). Early AWE systems focused primarily on essay scoring. Some of these systems adopt machine learning approaches that either target specific textual features, such as grammar or lexical sophistication, or focus on semantic similarity using techniques like latent semantic analysis. For instance, E-rater (Attali and Burstein, 2006) from ETS utilizes a suite of handcrafted linguistic features, such as grammar errors, vocabulary sophistication, organization indicators, etc. to evaluate writing. The Intelligent Essay Assessor (IEA) (Landauer et al., 2003) employs latent semantic analysis to measure the semantic similarity between an essay and high-scoring responses. More recent AWE systems have adopted deep learning techniques, in which models learn distributed representations of essays such that texts of similar quality are mapped to similar vector spaces (Li and Ng, 2024). One example is the work of Taghipour and Ng (2016), who employed a convolutional neural network (CNN) to extract n-gram-level features for capturing local dependencies, followed by a long short-term memory network (LSTM) to model global, long-distance dependencies for holistic essay scoring.

While some of the above systems have proven effective for scoring, such as E-rater (Burstein et al., 2004), a common limitation is their lack of true understanding on content, particularly higher-order thinking such as logic and argumentation. Thus, feedback in early AWE system, if any, tends to be formulaic (e.g., pointing out grammatical and mechanical errors) rather than giving insight into argument strength or coherence (Li and Ng, 2024).

2.2 LLM-based approaches

The recent rise of LLMs has sparked a new wave of research and applications in AWE. With their strong text comprehension and generation capabilities, LLMs can read a student essay and produce a detailed critique in natural language, often well beyond the templated feedback of older AWE systems. Researchers' efforts to utilize LLMs for AWE tasks include both direct prompting and fine-tuning approaches. In the realm of direct prompting, a study by Mansour et al. (2024) evaluated the capabilities of ChatGPT, the web interface of the GPT models, and LLaMA-2 in scoring written essays. Through various prompt-engineering tactics, they found that both models exhibited comparable performance in automated essay scoring, with ChatGPT having a slight advantage. Another study by Stahl et al. (2024) explored zero-shot and few-shot approaches inspired by Chain-of-Thought prompting (Wei et al., 2022) to generate both scores and feedback. Their study found that addressing both tasks simultaneously, rather than independently, enhances the quality of generated feedback and improves scoring performance, although the impact on feedback quality remains limited.

For studies that involve fine-tuning, researchers have found that in scoring, fine-tuning base models or even older or smaller models has been proven more effective than directly prompting more advanced LLMs (Li and Ng, 2024). A study by Wang and Gayed (2024) fine-tuned the GPT-3.5 model on a corpus of TOEFL argumentative essays with human benchmark scores and compared fine-tuned models with base GPT-3.5 and GPT-4 models. Their results demonstrated that the fine-tuned models achieved higher scoring accuracy and reliability than zero-shot prompting of GPT-3.5 or GPT-4, and that the fine-tuned models were robust when scoring essays from unseen prompts. Another study by Cai et al. (2025) introduced the Rank-Then-Score (RTS) framework, which employs a two-stage process: first ranking essays using a fine-tuned LLM, then scoring them based on the rankings. This method outperformed traditional supervised fine-tuning techniques, particularly in Chinese datasets. Further integrating fine-tuning and prompt engineering, Chu et al. (2024) proposed the Rationale-based Multiple Trait Scoring (RMTS) model. RMTS combines prompt-engineering-based LLMs with a fine-tuning-based essay scoring model to provide trait-specific rationales for scores. Their approach enhances the reliability of multi-trait scoring by generating fine-grained explanations aligned with rubric guidelines.

3 Methods

3.1 Dataset and subsets

The development of our system required a dataset of argumentative essays with benchmark scores and comprehensive feedback. To this end, we obtained a proprietary dataset of 480 TOEFL Independent Writing test-taker samples with official scores from ETS[3]. TOEFL Independent Writing is a typical argumentation writing task where test-takers write in response to an essay prompt, such as "*Do you agree or disagree the following statement:...*" The samples in the dataset are evenly distributed under two essays prompts and two ETS raters delivered integer scores from 0-5 based on specific rubrics[4]. Where the discrepancy of the two rater scores was no more than 1, the final score was the average of the two. Otherwise, a third rater was engaged to deliver the final score (Blanchard et al., 2013). As essays scored 0 were excluded, the scores of the essays in the dataset range from 1-5 with 0.5 increments. The score distribution of the 480 essays are shown in Table 7.

For the scoring module, we selected 120 essays through score-based equal sampling across the two essay prompts as the fine-tuning subset. As the number of essays scored 1 is limited, we added more essays of other scores to make up for it. The remaining 360 essays comprised the test subset.

For the feedback module, as no feedback was available in the dataset, we curated our own datasets for comprehensive feedback. For surface-level feedback in the form of corrective edits to grammar and mechanics, previous research on GEC has shown that directly prompting LLMs generates satisfactory results (Davis et al., 2024). As such, supervised fine-tuning and the creation of corresponding datasets were not necessary. For deep-level feedback that require higher-order thinking, we selected 90 essays from the 480 essays using score-based equal sampling. Eight experienced university teachers who teach academic writing courses were recruited to provide feedback using Microsoft (MS) Word's comment feature (see Figure 2 for an example of the comments). Before

[3] Access to this dataset is restricted to researchers approved by ETS, and therefore cannot be made publicly available. The dataset is provided under a limited, non-exclusive, revocable, and non-transferable license.

[4] https://www.nafsa.org/sites/default/files/ektron/files/underscore/regiii/conference/2009/pruner%20indepwrihd-t2.pdf

Module - Data Subset	Number of Essays	Data Content
Scoring - Fine-tune	120	Essay prompts, essays, scores, rubrics
Scoring - Test	360	Essay prompts, essays, scores, rubrics
Surface-level -Test	40	Essay prompts, essays
Deep-level - Fine-tune	90	Essay prompts, essays, target text elements, comments
Deep-level - Test	40	Essay prompts, essays (same with surface-level testing)

Table 1: Datasets used in fine-tuning and testing

annotation, the teachers received training on the scoring rubrics and were given nine benchmark essays scored from 1 to 5 and annotated by the authors as reference exemplars. The 90 essays had already been processed for surface-level corrections and the scores had been removed to ensure the annotations would focus solely on deep-level features. After the teachers submitted their annotated essays, another independent teacher reviewed all feedback comments to ensure they were accurate and free from surface-level errors. We then processed their feedback data into JSON format, including the targeted text elements (the start and end character positions of the text string and the tokens in the text element) and the respective comments. This JSON data comprises the fine-tuning subset for deep-level feedback.

For testing of both surface-level and deep-level feedback, we selected 40 essays (equal sampling) from the 390 non-annotated essays. The summary of datasets used in this study is shown in Table 1.

3.2 LLM choices and procedures

For the scoring module, we evaluated both open-source and commercial models to ensure broad applicability. For the open-source option, we selected LLaMA-3.3-70B-Instruct, the top model in instruction following[5]. For the commercial model, following suggestions in Wang and Gayed (2024), we chose GPT-4o, the latest model from the GPT series that allows supervised fine-tuning at the time this research took place. The fine-tuning prompt can be found at Appendix C.

For the surface-level feedback module that relies solely on direct prompting, we did a pilot experiment with SOTA LLMs including LLaMA 3.3-70B Instruct, GPT-4o and Claude 3.7 Sonnet using one essay for the GEC task. We found that GPT-4o had the best performance, followed by LLaMA, while Claude 3.7 exhibited over-correction, e.g., changing a word to its synonym of a higher register. Thus, we adopted GPT-4o and instructed the model to correct grammatical and mechanical errors in an essay and return a corrected version of the essay. Our prompt specifically asked it not to introduce stylistic and word choice corrections. The prompt can be found at Appendix D.

For the deep-level feedback module, supervised fine-tuning with our curated dataset of 90 essays relied on same models as the scoring module (LLaMA-3.3-70B and GPT-4o). For LLaMA, we fine-tuned the model using an entropy-based loss function over 8 epochs[6]. Meanwhile for GPT-4o, we relied on unknown default fine-tuning parameters due to the black box nature of the commercial model. For direct prompting, we aimed for multi-trait feedback inspired by Chu et al. (2024) and conducted thorough thematic analysis of the teacher feedback in the training subset. The resulting multiple traits, or structured category codes, are shown in Table 2. Based on the two primary categories of macro and micro feedback, we further divided the deep-level feedback into two pipelines. For model selection, we piloted GPT-4o, DeepSeek-R1, LLaMA-3.3-70B-Instruct and Claude 3.7 with one essay from the test subset. Among them, Claude 3.7 yielded the most satisfactory result and was subsequently adopted as the LLM in both pipelines. We then engineered long few-shot prompts to instruct Claude 3.7 to deliver feedback focusing on these subcategories and offer revision suggestions in both pipelines.

All source codes and prompts in this section can be found at xxxxx[7].

3.3 Validation scheme

3.3.1 Score validation

For the scoring module, we tested the fine-tuned LLaMA-3.3-70B-Instruct and GPT-4o models us-

[5] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

[6] The fine-tuned LLaMA model together with its detailed parameters can be found at [URL anynomized for review]

[7] Github link anynomized for review

Category	Subcategories
Macro feedback	Proper paragraphing; Text length appropriateness; Structure (introduction, body, and conclusion sections)
Micro feedback	Context-dependent grammar issues; Clarity of expression; Coherence between ideas; Argumentation quality and logical flow; Formality and academic register

Table 2: Feedback Categories for Deep-Level Analysis

ing the test subset of 360 essays and the same prompt as the fine-tuning prompt. To evaluate scoring accuracy and reliability, we employed three commonly used metrics: Root Mean Square Error (RMSE), Quadratic Weighted Kappa (QWK), and percentage agreement following Wang and Gayed (2024), whose results serve as a baseline in this study. RMSE is a measure of accuracy in the form of the absolute difference between system scores and benchmark scores, with lower values indicating better performance (Chai and Draxler, 2014). QWK assessed the level of agreement between system and benchmark scores while considering the ordinal nature of score categories (Li and Ng, 2024). Percentage agreement is used as a complement to QWK, and we measured the proportion of essays of exact agreement (same score) and adjacent agreement (a difference of 0.5) with benchmark scores.

3.3.2 Feedback validation

For surface-level feedback, we aligned each original essay and its corresponding corrected version using ERRANT v3.0.0[8] (ERRor ANnotation Toolkit; Bryant et al., 2017), a grammar error annotation tool that automatically detects and categorizes all edit operations. Two expert raters evaluated each identified edit along two dimensions:

- **Necessity**: Whether the identified text element genuinely requires correction, given the reported tendency of LLMs to over-correct (Katinskaia and Yangarber, 2024).
- **Effectiveness**: Whether the suggested corrective edit effectively addresses the error.

The two raters discussed each edit and provided additional comments for cases where edits were deemed unnecessary or ineffective.

For deep-level feedback, we first performed a preliminary check on the integrity of LLM output and found that the two fine-tuned models were not competent for the task. For the fine-tuned GPT-4o model, it produced incomplete output for every inference. For each essay, it generated so many comments that the output token exceeded the context window of 8,000 tokens. Thus, each inference was terminated before the complete output was generated. Meanwhile, the fine-tuned LLaMA-3.3-70B model produced formatting errors that prevent its output from being parsed as valid JSON. The most severe issues were structure errors, such as missing key-value delimiters (`{"highlighted", "data":"..."}` instead of `{"highlighted": "...", "data":"..."}`) and unclosed braces (`[{"data": "...", {"data":"..."}]` instead of `[{"data": "..."}, {"data":"..."}]`). On the other hand, Claude 3.7 successfully generated a sufficient number of feedback comments in the correct data format through direct prompting. As such, we decided to only perform human evaluation on valid output from Claude 3.7. In our prompt to the macro feedback pipeline, we asked Claude 3.7 specifically to give feedback to each paragraph, thus the two raters only assessed whether each macro feedback comment was *effective*. For micro feedback, each was assessed on the dimensions of *necessity* and *effectiveness*, similar to surface-level feedback evaluation. Following the independent evaluations, we calculated inter-rater reliability. For any cases where the two primary raters disagreed, a third expert rater was consulted to deliver the final judgment.

4 Results

4.1 Scoring module

The RMSE, QWK and percentage agreement between the essay scores generated by the two fine-tuned models and benchmark scores are shown in Table 3. The baseline model is the best performing fine-tuned model in Wang and Gayed (2024), the only prior study that used the same dataset for essay scoring.

Results indicate SOTA performance of our two

[8] https://github.com/chrisjbryant/errant

Fine-tuned Model	RMSE	QWK	Percent (absolute)	Percent (adjacent)	Percent (total)
GPT-4o	0.44	0.84	0.45	0.48	0.93
LLaMA	0.53	0.81	0.41	0.45	0.86
Baseline (Wang and Gayed, 2024)	0.57	0.78	0.33	0.52	0.85

Table 3: Performance metrics of fine-tuned models on the test subset

fine-tuned models in scoring accuracy and reliability, with GPT-4o consistently performing the best across all three evaluation metrics. Specifically, both our fine-tuned models achieved QWK scores above 0.8, surpassing the commonly accepted threshold for near-perfect agreement (Sim and Wright, 2005). For context, ETS considers a QWK of 0.7 to be sufficient for reliable scoring of TOEFL Independent Writing tasks by its E-rater system (Williamson et al., 2012); both of our fine-tuned models exceeded this benchmark. Although ETS does not specify a formal threshold for RMSE, the observed discrepancies of 0.44 (GPT-4o) and 0.53 (LLaMA) from benchmark scores are reasonable, given that ETS permits up to a one-point difference between two human raters (Blanchard et al., 2013), which results in a final averaged score deviating by 0.5 from each individual score.

4.2 Surface-level feedback module

After the 40 essays in the test subset were corrected in the surface-level module, ERRANT tagged 2049 edit operations. Human evaluation deemed 1985 edits deemed necessary (96.88%), out of which 1970 were deemed both necessary and effective (96.14%). Among the unnecessary edits, two salient patterns have been identified from rater comments. First, GPT-4o seems to prefer British use than American use (N=7). For example, it changed *favorite* to *favourite* and moved a period or a comma outside a closing quotation mark (e.g., “*...store.*” to “*...store*”.) to match the British style. The second pattern is related to comma additions (N=15). In particular, GPT-4o preferred adding the Oxford comma before the last item in a list of nouns (N=10; e.g., *A, B and C* into *A, B, and C*), which was deemed unnecessary by raters.

4.3 Deep-level feedback module

4.3.1 Macro feedback pipeline

The 40 essays in the test subset contained 201 paragraphs and thus Claude 3.7 generated 201 macro feedback comments. The contingency table between the two raters on the effectiveness of the comments is shown in Table 4

Rater A / Rater B	Effective	Ineffective
Effective	179	9
Ineffective	13	0

Table 4: Contingency table of macro comment evaluation by two raters

To assess inter-rater reliability, we employed Gwet’s $AC1$ coefficient (Gwet, 2008), which ranges from -1 to $+1$ similar to Cohen’s Kappa. This measure was selected over Cohen’s Kappa due to its robustness against the “Kappa paradox”, where high observed agreement coincides with low kappa values in cases of skewed marginal distributions (i.e., the dominance of the *effective* category in rater evaluation) (Wongpakaran et al., 2013; Gwet, 2008). The $AC1$ coefficient was calculated to be 0.89, with a standard error (SE) of 0.03 and a 95% confidence interval (CI) ranging from 0.82 to 0.93, indicating very strong[9] and statistically significant agreement beyond chance ($p < .001$).

After a third rater resolved the disagreement of the two raters, the number of effective comments was finalized as 187 (93.03%) and that of ineffective comments, 27 (6.97%). For the 27 ineffective comments, remarks from raters revealed that Claude 3.7 was not flexible enough to handle unanticipated input variations. For example, one test-taker included a title in their essay, a component not required in the writing task, and Claude 3.7 misinterpreted it as the first paragraph. In another case, some test-takers improperly formatted each sentence as a separate paragraph, which led to repeated comments that a paragraph was too short and thus should be combine with the previous one (N=8). In addition, some test-takers were unable to complete the task within the allot-

[9] As there is no commonly accepted interpretation of Gwet’s $AC1$ coefficient, we relied on the interpretation of Cohen’s Kappa to determine the degree of agreement (Landis and Koch, 1977)

Rater A / Rater B	Necessary	Unnecessary
Necessary	579	28
Unnecessary	22	1

Table 5: Contingency table of necessity of micro feedback by two raters

Rater A / Rater B	Effective	Ineffective
Effective	577	1
Ineffective	1	0

Table 6: Contingency table of effectiveness judgment for micro feedback by two raters

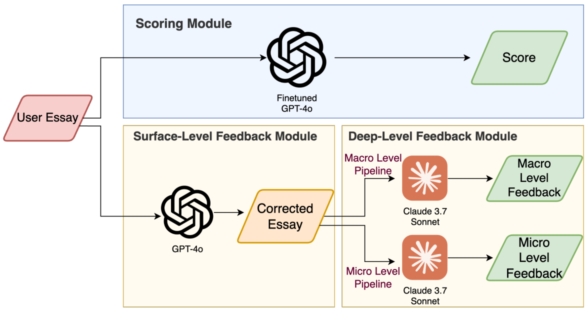


Figure 1: Final system architecture

ted time. Claude 3.7 failed to recognize the time constraint of such unfinished responses and instead provided detailed guidance on composing conclusions (N=2). Also very interesting is that Claude 3.7 was not aware of the test-taker situation where they didn't have access to external resources and recommended incorporating research evidence and citations to strengthen arguments (N=1).

4.3.2 Micro feedback pipeline

For the same 40 essays, Claude 3.7 generated 630 micro feedback comments. The contingency table between the two raters on the necessity of the comments is show in Table 5, while that on the effectiveness of necessary comments mutually agreed by two raters is shown in Table 6.

Given the skewed distribution of judgments in both dimensions, we again employed Gwet's $AC1$ (Gwet, 2008). The resulting agreement on necessity was very strong and statistically significant beyond chance ($AC1$= 0.91; SE=0.01; 95% CI, [0.89, 0.94]; $p < .001$). The resulting agreement on effectiveness was also very strong and statistically significant beyond chance ($AC1$= 1.00; SE=0.00; 95% CI, [0.99, 1.00]; $p < .001$).

After a third rater resolved the disagreement between the two raters, the number of necessary comments was finalized as 600 (95.24%), that of unnecessary ones, 30 (4.76%). Among the necessary comments, 596 were deemed effective while 4 were ineffective, and thus the proportion of both necessary and effective comments is 94.69%.

Based on rater comments, one salient pattern was observed: some micro comments overlapped with macro ones. Claude 3.7 at times expand the comment for a sentence to cover the whole paragraph (N=18). For instance, when the first sentence in an introduction paragraph did not properly set the background of the topic, Claude 3.7 would comment that there was a lack of proper background and then went on to give more feedback on how to write a good introduction paragraph.

4.4 Final system architecture and UI desgin

Based on the evaluation results, we finalized the system architecture by selecting LLMs tailored to each module: the fine-tuned GPT-4o model for the scoring module; direct prompting with GPT-4o for the surface-level feedback module; and Claude 3.7 for both pipelines in the deep-level feedback module. Figure 1 illustrates the overall workflow of our system.

To mimick human feedback mechanisms, we developed an interactive UI that consists of a user input page and a feedback page[10]. The user input page is where an user provides the essay prompt and their writing (see Figure 3 in Appendix E). The feedback page shows the predicted score of an essay and includes a surface-level feedback tab and a deep-level feedback tab. The surface-level feedback tab allows three display modes of surface-level feedback: (1) the original writing with erroneous elements highlighted, (2) a combination of the original and corrected writing with edit operations shown, and (3) the corrected writing with corrections highlighted (see Figure 4 in Appendix E). For the deep-level feedback tab, we designed a comment-based interface modeled after Microsoft Word's Comment feature. Macro-level comments are displayed in a left-side panel and are aligned with paragraph boundaries. Micro issues under coherence, clarity, grammar, argumentation and formality are highlighted within the essay text using color-coded categories. When a user hovers the mouse over a highlighted section, the correspond-

[10] The web UI can be accessed through [website anynomized for review].

ing feedback comment appears on the right-side panel (see Figure 5 in Appendix E).

5 Discussions

Our system built on a modular approach successfully delivered accurate and reliable scores benchmarked against ETS standards for TOEFL independent writing and high-quality feedback validated by human teachers. The findings further support previous research by Stahl et al. (2024) and Chu et al. (2024), who have demonstrated that specialized models handling different aspects of writing evaluation can produce better assessments than unified approaches.

Human evaluation of surface-level feedback shows that GEC tasks can be addressed effectively by well-designed prompts without the need for sophisticated fine-tuning. This finding echoes the study by Zeng et al. (2024), who demonstrated that LLMs with appropriate prompting can achieve competitive performance on GEC tasks compared to specialized fine-tuned models. Particularly, we successfully minimized over-correction through our targeted prompt, an issue previously identified as a key limitation of LLM-prompting-based GEC approach (Katinskaia and Yangarber, 2024; Davis et al., 2024). The high precision rate of 96.14% indicates that when the system identifies errors, it is nearly always correct in doing so.

Human validation of deep-level feedback suggests that well-structured few-shot prompting can generate effective and structured comments on higher-order thinking, and may even be more effective than models of supervised fine-tuning. Here we would like to focus on why supervised fine-tuning did not yield satisfactory results on deep-level feedback. The fine-tuned GPT model that provided feedback to almost every word, phrase, and sentence, most of which were unnecessary. As the fine-tuning parameters were unknown to us, we can only speculate that supervised fine-tuning with non-deterministic output is not well-suited for GPT-4o. Contrary to numerical and categorical output, there are no clear standards as to whether a comment is necessary and/or effective. Further reinforcement learning from human feedback (RLHF) may be needed to guide the fine-tuned GPT-4o to understand what feedback is necessary and preferred. For the fine-tuned LLaMA model, despite the data parsing issues, we found that the feedback comments were actually quite similar to those in the fine-tuning subset, though still not so good as those from Claude 3.7. This indicates that the supervised fine-tuning did enhance model performance in generating feedback. We believe the data formatting issue is likely an inherent issue with LLaMA, as GPT-4o did not exhibit such an issue. A previous study by Yao et al. (2025) specifically highlighted the limitations of LLaMA in data formatting tasks, which prompted the researchers to develop a format benchmark to evaluate and improve formatting capabilities of LLMs. If the data formatting issue can be addressed, perhaps by incorporating a data checker tool, it may be possible to effectively use fine-tuned LLaMA and potentially other open-source models for generating structured feedback.

With an interactive UI, our system has strong pedagogical implications. The system can immediately benefit students engaged in argumentative writing. They can receive prompt feedback not available in traditional classroom settings, where it usually takes teachers days to provide comprehensive feedback on student writing. Teachers can also use this tool to alleviate their workload, particularly for surface-level feedback. Instead of spending time correcting basic grammatical and mechanical errors, they can focus more on guiding students' higher-order thinking in argumentative writing, aided by the feedback generated by the system. At the institutional level, the system can also serve as a reference for scoring to ensure fairness. Particularly for large-scale coordinated writing courses taught by various instructors, it is common for teachers to develop idiosyncratic scoring criteria. Having an automated reference score could promoting more consistent evaluation standards.

6 Conclusions

In this study, we successfully built *WrAFT*, a modular LLM-powered automated writing evaluation system designed specifically for argumentative essays. By decomposing AWE tasks into scoring, surface-level feedback, and deep-level feedback modules, we achieved state-of-the-art performance in scoring on a proprietary TOEFL writing dataset. Human validation of system-generated feedback also suggest satisfactory results. Our findings provide a strong foundation for future research in modular AWE systems that provide both scoring and comprehensive feedback. An interactive UI has also been developed for public use free of charge.

Limitations

There are a few limitations in the design of this study. First, we developed our system using a proprietary dataset that has only been used in one prior AWE study. This limits the scope of direct comparisons on scoring performance. Second, in our validation scheme, a key limitation is that human evaluation only focused on precision without considering recall. For both surface-level and deep-level feedback, we assessed the necessity and effectiveness of feedback but did not evaluate whether the system captured all errors that should have been identified or all text elements that should have required feedback comments. A more robust evaluation would require establishing a gold standard of all possible errors and feedback opportunities against which system performance could be measured. Unfortunately, the extensive human resources required to achieve this was beyond our capacity. In addition, there is a lack of measurement of actual impact on student learning outcomes, which was beyond the scope of this research. Future research should include experimental studies tracking writing improvement over time when using the system compared to traditional feedback methods.

There are limitations with the system as well. First, apart from the salient issues identified through rater comments, we also observed instances of LLM hallucinations. For example, there were cases, though infrequent, where the system identified non-existent errors and subsequently suggested "corrections" that were identical to the original text. Future iterations should incorporate verification mechanisms, such as adding another LLM as the reviewer of generated feedback. Second, the system is limited in its generalizability, as it was developed and tested exclusively on argumentative essays with writing prompts. It cannot be directly applied to other types of argumentative writing, such as source-based writing where writers are required to incorporate or respond to source information. Third, the feedback offered by the system is solely in English, creating barriers for learners who might struggle to comprehend feedback comments. Future iterations should incorporate multilingual feedback capabilities, including translation of comments into learners' first languages to facilitate comprehension and implementation of suggestions.

From a practical implementation perspective, there are limitations arising from the reliance on commercial API services. There are potential scalability issues related to cost, rate limits, and long-term sustainability. In addition, constant changes to commercial models may impact system performance. Data privacy concerns also emerge when student writing is processed through third-party services.

Ethics Statement

This research was conducted in compliance with ethical standards for working with human data and received approval from the Institutional Review Board (IRB) at [Anonymized University], protocol number 2024-305. All essays in the dataset were fully anonymized and contain no personally identifiable information.

All annotators participated voluntarily and were financially compensated at a standard hourly rate in accordance with institutional guidelines. Annotator identities were anonymized in all records.

The WrAFT system relies on commercial LLM APIs (GPT-4o, Claude 3.7) for model inference. While only de-identified text was submitted to these APIs during development and testing, we acknowledge potential privacy concerns in real-world deployment scenarios and encourage practitioners to evaluate data governance policies before use in educational settings.

Finally, this work is intended to support educators and learners by augmenting, not replacing, human judgment in writing assessment. The system is designed for formative, instructional use and should not be used in high-stakes evaluation contexts without appropriate oversight.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3).

John Bitchener and Ute Knoch. 2008. The value of written corrective feedback for migrant and international students. *Language Teaching Research*, 12(3):409–431.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013:i–15.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *Ai magazine*, 25(3):27–27.

Yida Cai, Kun Liang, Sanwoo Lee, Qinghan Wang, and Yunfang Wu. 2025. Rank-then-score: Enhancing large language models for automated essay scoring. *arXiv preprint arXiv:2504.05736*.

T. Chai and R. R. Draxler. 2014. Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. 2024. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. *arXiv preprint arXiv:2410.14202*.

Christopher Davis, Andrew Caines, Øistein E. Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of English learner text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11952–11967, Bangkok, Thailand. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Gerald Graff. 2003. *Clueless in Academe: How Schooling Obscures the Life of the Mind*. Yale University Press.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2019. Minimally-augmented grammatical error correction. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 357–363, Hong Kong, China. Association for Computational Linguistics.

Kilem L Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. Grammar error correction using pseudo-error sentences and domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 388–392, Jeju Island, Korea. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097, Melbourne, Australia. Association for Computational Linguistics.

Anisia Katinskaia and Roman Yangarber. 2024. GPT-3.5 for grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.

Deanna Kuhn. 2005. *Education for Thinking*. Harvard University Press.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. In *Automated Essay Scoring: A Cross-disciplinary Perspective*, pages 87–112. Lawrence Erlbaum Associates.

J. Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jianwei Li and Jiahui Wu. 2023. Automated essay scoring incorporating multi-level semantic features. In *International Conference on Artificial Intelligence in Education*, pages 206–211. Springer.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.

Sha Liu and Antony Kunnan. 2015. Investigating the application of automated writing evaluation to chinese undergraduate english majors: A case study of writetolearn. *Assessing Writing*, 24:1–15.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? *arXiv preprint arXiv:2403.06149*.

Ellis B Page. 1967. Grading essays by computer: Progress report. In *Proceedings of the invitational Conference on Testing Problems*.

Ellis Batten Page. 2003. *Project Essay Grade: PEG*. Lawrence Erlbaum Associates Publishers.

Jim Ranalli and Junko Yamashita. 2021. L2 student engagement with automated feedback on writing. *System*, 99:102512.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Mark D. Shermis and Jill C. Burstein. 2003. *Automated Essay Scoring: A Cross-disciplinary Perspective*. Lawrence Erlbaum Associates.

Mark D Shermis and Joshua Wilson. 2024. Introduction to automated essay evaluation. In *The Routledge international handbook of automated essay evaluation*, pages 3–22. Routledge.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 234–245.

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*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

John Truscott. 1996. The case against grammar correction in l2 writing classes. *Language Learning*, 46(2):327–369.

Ivo Verhoeven, Pushkar Mishra, Rahel Beloch, Helen Yannakoudakis, and Ekaterina Shutova. 2024. A (more) realistic evaluation setup for generalisation of community models on malicious content detection. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 437–463, Mexico City, Mexico. Association for Computational Linguistics.

J. Wang and M. Gayed. 2024. Effectiveness of large language models in automated evaluation of argumentative essays. *Computer Assisted Language Learning*, 37(1):1–25.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13(1):61.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Jiashu Yao, Heyan Huang, Zeming Liu, Haoyu Wen, Wei Su, Boao Qian, and Yuhang Guo. 2025. Reff: Reinforcing format faithfulness in language models across varied tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25660–25668.

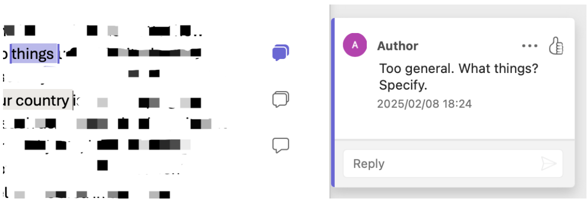
Min Zeng, Jiexin Kuang, Mengyang Qiu, Jayoung Song, and Jungyeul Park. 2024. Evaluating prompting strategies for grammatical error correction based on language proficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6426–6430, Torino, Italia. ELRA and ICCL.

A Score distribution in the dataset

Score	Prompt 1	Prompt 2	Subtotal
1	1	2	3
1.5	3	2	5
2	18	20	38
2.5	32	25	57
3	65	55	120
3.5	40	45	85
4	32	38	70
4.5	35	28	63
5	14	25	39
Total	240	240	480

Table 7: Score distribution in the dataset

B Example of a teacher comment in MS Word



[redacted]

things

ur country

Author

Too general. What things? Specify.

2025/02/08 18:24

Figure 2: Example of a teacher comment in MS Word

C Fine-tuning and inference prompt for the scoring module

```
<instructions>
As a language expert, your task is to evaluate
argumentative essays on a scale of 0 to 5
```

```
(with 0.5 increments) based on the rubrics below.

<rubric>
[Complete rubrics here]
</rubric>
</instructions>

<input_data_structure>
{{
  ”essay_prompt”: ”...”, // The essay prompt
  ”essay_text”: ”...”, // The essay text
}}
</input_data_structure>

<output_data_structure>
{{
  ”score”: 0.5 // The score
}}
</output_data_structure>

Here's the input data:
<input_data>
{{
  ”essay_prompt”: {essay_prompt},
  ”essay_text”: {essay_text}
}}
</input_data>
```

D Surface-level prompt

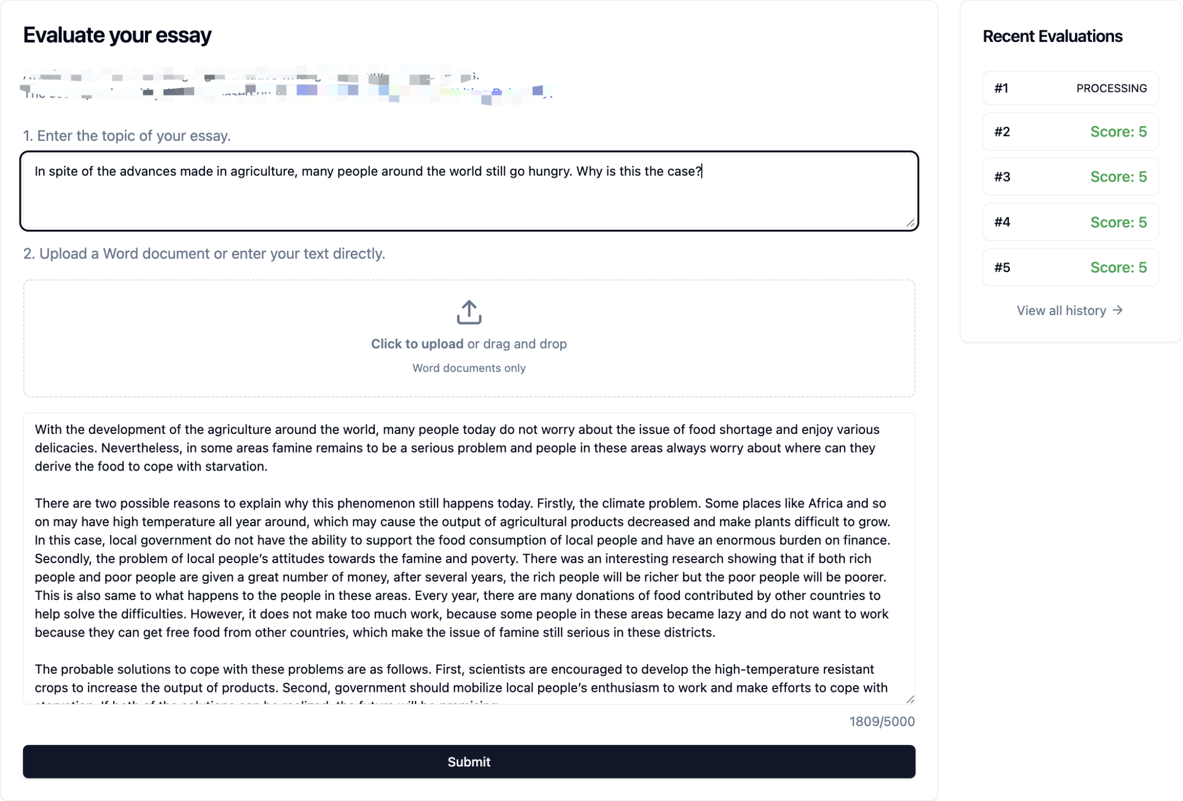
```
<instructions>
You are an English linguist and your task is to
correct the grammatical and mechanical errors in
an English essay. Do not alter word choices
unnecessarily (e.g., replacing words with synonyms)
or make stylistic improvements. Keep the original
paragraphing and DO NOT remove or add any paragraphs.

Requirements:
1. The output should be in JSON format.
2. The output should be in the same format
as the output_data_structure.
</instructions>

<output_data_structure>
{{
  ”corrected_paragraphs”: [
    ”...”, // The corrected text of the first paragraph
    ”...”, // The corrected text of the second paragraph
    ...
  ]
}}
</output_data_structure>

Here's the input data:
<input_data>
{{
  ”essay_paragraphs”: {essay_paragraphs}
}}
</input_data>
```

E UI screenshots



Evaluate your essay

[redacted]

1. Enter the topic of your essay.

In spite of the advances made in agriculture, many people around the world still go hungry. Why is this the case?

2. Upload a Word document or enter your text directly.

With the development of the agriculture around the world, many people today do not worry about the issue of food shortage and enjoy various delicacies. Nevertheless, in some areas famine remains to be a serious problem and people in these areas always worry about where can they derive the food to cope with starvation.

There are two possible reasons to explain why this phenomenon still happens today. Firstly, the climate problem. Some places like Africa and so on may have high temperature all year around, which may cause the output of agricultural products decreased and make plants difficult to grow. In this case, local government do not have the ability to support the food consumption of local people and have an enormous burden on finance. Secondly, the problem of local people's attitudes towards the famine and poverty. There was an interesting research showing that if both rich people and poor people are given a great number of money, after several years, the rich people will be richer but the poor people will be poorer. This is also same to what happens to the people in these areas. Every year, there are many donations of food contributed by other countries to help solve the difficulties. However, it does not make too much work, because some people in these areas became lazy and do not want to work because they can get free food from other countries, which make the issue of famine still serious in these districts.

The probable solutions to cope with these problems are as follows. First, scientists are encouraged to develop the high-temperature resistant crops to increase the output of products. Second, government should mobilize local people's enthusiasm to work and make efforts to cope with

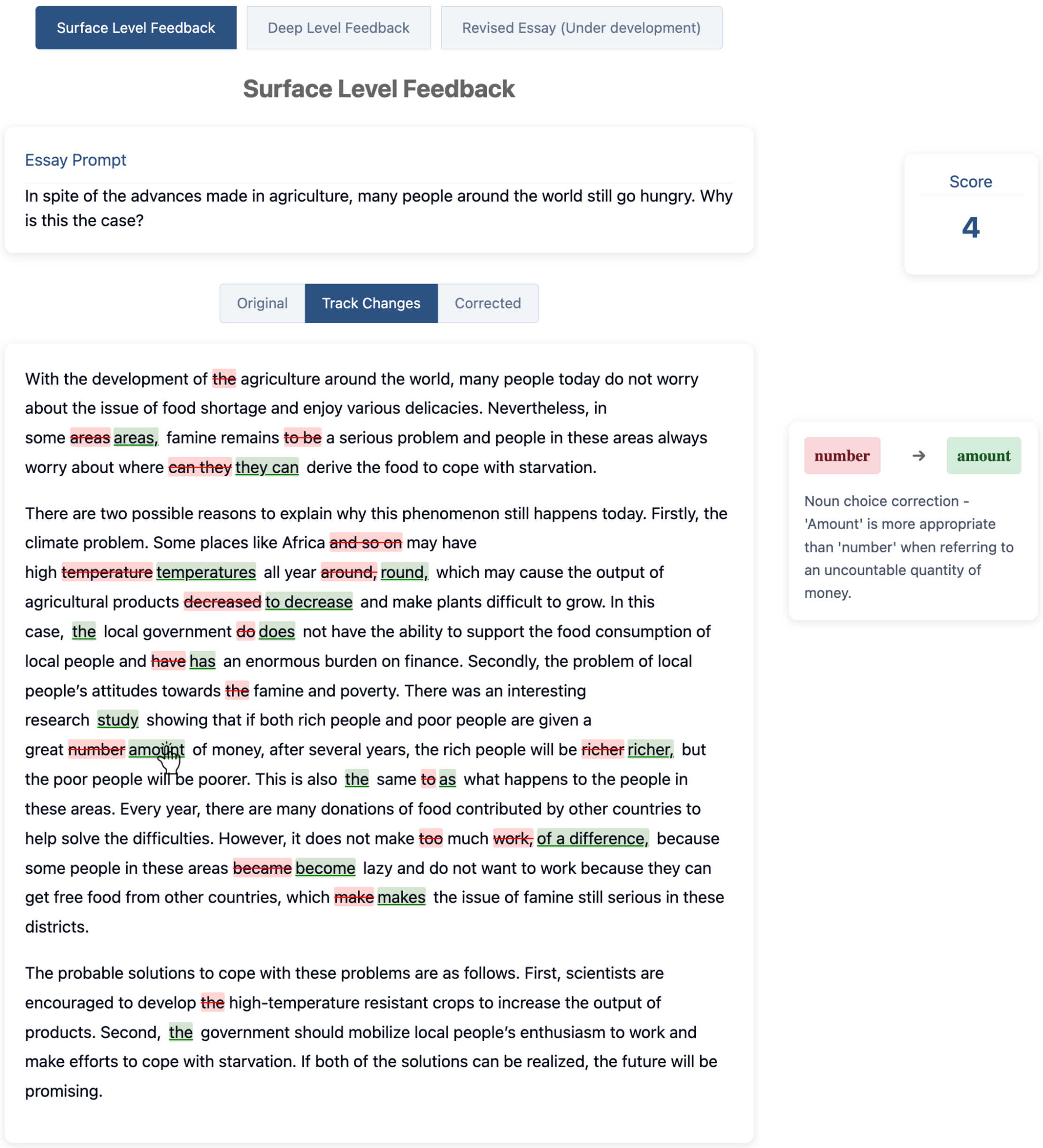
1809/5000

Recent Evaluations

#1	PROCESSING
#2	Score: 5
#3	Score: 5
#4	Score: 5
#5	Score: 5

Figure 3: Screenshot of the user input page

Note: A user can input the essay prompt and their writing (through either uploading a Word document or pasting the text directly) in the interface. A list of recent evaluations is shown on the right side bar. The sample essay in the screenshot is sourced from a Chinese university student, independent from the TOEFL writing dataset used in this study. The explanations under "Evaluate your essay" are blurred out for anonymity in peer review.



Surface Level Feedback

Essay Prompt

In spite of the advances made in agriculture, many people around the world still go hungry. Why is this the case?

Score

4

With the development of ~~the~~ agriculture around the world, many people today do not worry about the issue of food shortage and enjoy various delicacies. Nevertheless, in some ~~areas~~ areas, famine remains ~~to be~~ a serious problem and people in these areas always worry about where ~~can they~~ they can derive the food to cope with starvation.

There are two possible reasons to explain why this phenomenon still happens today. Firstly, the climate problem. Some places like Africa ~~and so on~~ may have high ~~temperature~~ temperatures all year ~~around,~~ round, which may cause the output of agricultural products ~~decreased~~ to decrease and make plants difficult to grow. In this case, the local government ~~do~~ does not have the ability to support the food consumption of local people and ~~have~~ has an enormous burden on finance. Secondly, the problem of local people's attitudes towards ~~the~~ famine and poverty. There was an interesting research study showing that if both rich people and poor people are given a great ~~number~~ amount of money, after several years, the rich people will be ~~richer~~ richer, but the poor people will be poorer. This is also the same ~~to~~ as what happens to the people in these areas. Every year, there are many donations of food contributed by other countries to help solve the difficulties. However, it does not make ~~too~~ much ~~work,~~ of a difference, because some people in these areas ~~became~~ become lazy and do not want to work because they can get free food from other countries, which ~~make~~ makes the issue of famine still serious in these districts.

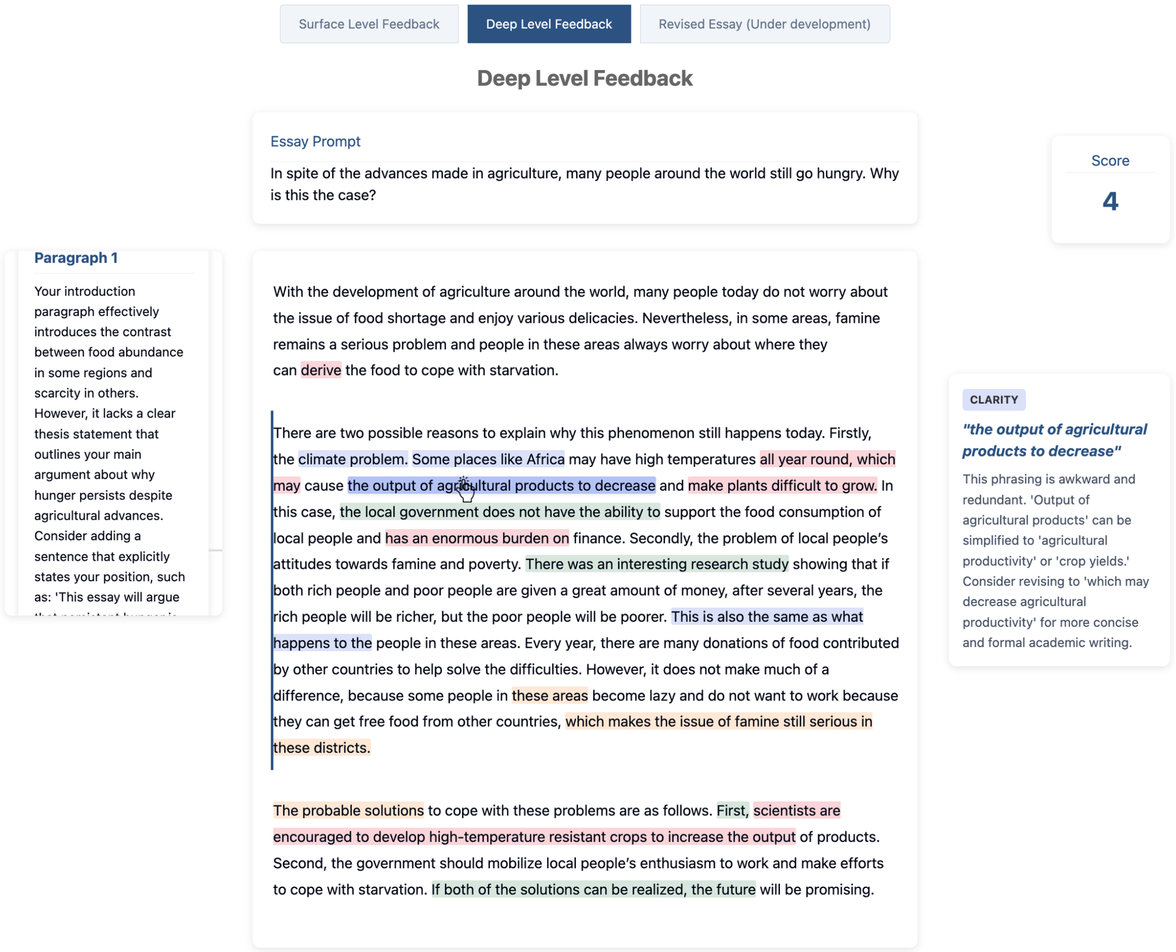
The probable solutions to cope with these problems are as follows. First, scientists are encouraged to develop ~~the~~ high-temperature resistant crops to increase the output of products. Second, the government should mobilize local people's enthusiasm to work and make efforts to cope with starvation. If both of the solutions can be realized, the future will be promising.

number → amount

Noun choice correction - 'Amount' is more appropriate than 'number' when referring to an uncountable quantity of money.

Figure 4: Screenshot of the surface-level feedback page - track changes

Note: The surface-level feedback can be displayed in three modes and the figure shows the track changes mode showing each edit operation. The short explanation shown when the mouse hovers upon a certain edit is an idea for future work not included in the present study.



Deep Level Feedback

Essay Prompt

In spite of the advances made in agriculture, many people around the world still go hungry. Why is this the case?

Score

4

Paragraph 1

Your introduction paragraph effectively introduces the contrast between food abundance in some regions and scarcity in others. However, it lacks a clear thesis statement that outlines your main argument about why hunger persists despite agricultural advances. Consider adding a sentence that explicitly states your position, such as: 'This essay will argue

With the development of agriculture around the world, many people today do not worry about the issue of food shortage and enjoy various delicacies. Nevertheless, in some areas, famine remains a serious problem and people in these areas always worry about where they can derive the food to cope with starvation.

There are two possible reasons to explain why this phenomenon still happens today. Firstly, the climate problem. Some places like Africa may have high temperatures all year round, which may cause the output of agricultural products to decrease and make plants difficult to grow. In this case, the local government does not have the ability to support the food consumption of local people and has an enormous burden on finance. Secondly, the problem of local people's attitudes towards famine and poverty. There was an interesting research study showing that if both rich people and poor people are given a great amount of money, after several years, the rich people will be richer, but the poor people will be poorer. This is also the same as what happens to the people in these areas. Every year, there are many donations of food contributed by other countries to help solve the difficulties. However, it does not make much of a difference, because some people in these areas become lazy and do not want to work because they can get free food from other countries, which makes the issue of famine still serious in these districts.

The probable solutions to cope with these problems are as follows. First, scientists are encouraged to develop high-temperature resistant crops to increase the output of products. Second, the government should mobilize local people's enthusiasm to work and make efforts to cope with starvation. If both of the solutions can be realized, the future will be promising.

CLARITY

"the output of agricultural products to decrease"

This phrasing is awkward and redundant. 'Output of agricultural products' can be simplified to 'agricultural productivity' or 'crop yields.' Consider revising to 'which may decrease agricultural productivity' for more concise and formal academic writing.

Figure 5: Screenshot of the deep-level feedback page

Note: The deep-level feedback page displays feedback comment to each paragraph on the left side bar and micro feedback on the right side bar once the mouse hovers upon a specific text element. Highlights in different colors represents the multiple traits of feedback. The page also shows a tab for "Revised Essay" (LLM-revised essay based on deep-level feedback), which is a new feature under development not included in the present work.