WrAFT: A Modularized Automated Writing Evaluation System for L2 Argumentative Essays

Adnan Labib[1], Yixuan Huang[2], Jiahui Wu[3], John Maurice Gayed[2], Zheng Yuan[4], Qiao Wang[5]
[1]Department of Informatics, King's College London, UK
[2]Waseda University, Tokyo, Japan
[3]Beijing University of Posts and Telecommunications, Beijing, China
[4]University of Sheffield, Sheffield, UK
[5]Hosei University, Tokyo, Japan
adnan.1.labib@kcl.ac.uk, yixuan.huang@moegi.waseda.jp, amy_wjh@bupt.edu.cn
gayed@waseda.jp, zheng.yuan1@sheffield.ac.uk, judy.wang@hosei.ac.jp

Abstract

Second language writers rarely receive the detailed feedback they need to improve, as comprehensive feedback takes teachers days per class. To solve this, we present *WrAFT*, a **Wr**iting **A**ssessment and **F**eedback **T**ool, that delivers reliable scores along with effective comprehensive feedback for argumentative essays. *WrAFT* adopts a modular design by dividing the automated writing evaluation (AWE) into essay scoring, surface-level feedback and deep-level feedback. In building the system, various Large Language Models (LLMs) have been evaluated, including LLaMA-3.3-70B-Instruct, GPT-4o, and Claude 3.7, through both supervised fine-tuning and direct prompting. Evaluation shows that *WrAFT* achieves strong agreement with human raters, with a Quadratic Weighted Kappa (QWK) of 0.84 on a proprietary dataset of 480 TOEFL Independent Writing essays. More importantly, human evaluation of system-generated feedback also reveals high approval ratings (96.14% for surface-level, 93.03% for deep-level macro feedback, and 94.69% for deep-level micro feedback). Publicly available interactive user interface has also been developed for the system.

1 Introduction

Second language writers need comprehensive feedback to improve, but rarely receive it as teachers may take days to provide feedback, and what learners get is often generic (Liu and Kunnan, 2015; Ranalli and Yamashita, 2021). 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, 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 the GRE and TOEFL writing exams (Attali and Burstein, 2006), they often failed to capture higher-order thinking in writing such as coherence or argumentation.

With the advancement of Large Language Models (LLMs), LLM-based AWE systems can now predict a score along with giving rich explanatory feedback. However, 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, 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

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

Proceedings of the 1st Workshop on Multi-modality and Multi-lingual Language Learning, pages 1–18
November 9, 2026 ©2026 Association for Computational Linguistics

caution against the potential cognitive overload that comprehensive feedback might impose on learners (Truscott, 1996), it still remains a prevalent pedagogical practice in language education, as learners may selectively attend to aspects they find most relevant. However, though there have been many grammatical error correction (GEC) systems that can be used for surface-level feedback (Yuan and Briscoe, 2016; Imamura et al., 2012; Rozovskaya and Roth, 2019; Davis et al., 2024; Bryant et al., 2023; Yuan et al., 2021), deep-level feedback remains rare and is usually in a generic manner. 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.

One notable AWE system, AiAWE (Gayed, 2025), provides highly reliable and accurate TOEFL-benchmarked scoring and holistic feedback to argumentative essays through LLMs fine-tuned on a proprietary TOEFL independent writing dataset provided by Education Testing Service (ETS). Inspired by the AiAWE system and drawing on the first two layers of feedback proposed by Li and Ng (2024), this research aims to address the current research gap by developing an LLM-based AWE system for argumentative essays that provides accurate scoring and layered, comprehensive feedback. Internally, a modular architecture is adopted with separate modules for scoring, surface-level feedback and deep-level feedback. This system is made available through an interactive user interface which transforms the system from a research model into a practical and pedagogically effective educational tool. The contributions are as follows:

- The scoring module achieves strong agreement with human raters (QWK of 0.84 on 1-5 score scale), allowing learners to understand their essay scores on the TOEFL rubric level.
- Human validation of the comprehensive feedback generated by the AWE 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.
- For feedback generation, which is 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.
- The research results in a deployable code package[2] and a web-based application to visualize comprehensive feedback through in-line corrective edits and anchored comments[3].

2 Background

2.1 Early automated writing evaluation

AWE is defined as "the process of evaluating and scoring written prose via computer programs" (Shermis and Wilson, 2024), which includes the dual task of scoring and feedback. Page (1967) first proposed computer based essay grading and developed Project Essay Grade (Page, 2003). Early AWE systems adopted machine learning approaches, 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. Another AWE system, 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 (Li and Ng, 2024), such as 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 essay scoring. While they emphasized the operational potential of these systems, later critiques (see Perelman, 2014) show that many of these scoring engines disproportionately reward surface-level features, especially essay length (word count), rather than substantive content, raising questions about the depth of construct representation.

Further investigation shows that while some of the above systems have proven effective for scoring, they lack true understanding of content, particularly higher-order thinking such as logic and argumentation. Thus, feedback in early AWE systems, 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

[2] https://github.com/judywq/wraft
[3] https://wraft.net/

Ng, 2024). In addition, while *e-rater* has been used extensively in high-stakes environments, Powers et al. (2002) demonstrates that writing experts and computational linguists were able to deliberately create essays that exploit the *e-rater* algorithm. Their findings highlight that automated systems could reward superficial linguistic features rather than the deeper qualities of coherent and original writing. Deane (2013) extends on this by arguing that while AWE systems can capture certain linguistic and structural features, they risk underrepresenting higher-order writing dimensions such as idea development and rhetorical effectiveness.

2.2 LLM-based approaches

Recently, LLMs, with their strong text comprehension and generation capabilities, can read a student essay and produce a detailed critique in natural language, often well beyond the templated feedback of older AWE systems. Mizumoto and Eguchi (2023) empirically evaluated the use of LLMs for automated essay scoring by using GPT-3 to assess essays from ETS's TOEFL11 dataset. The researchers reported the model achieving good levels of agreement and reliability with the human ratings by professional ETS raters. In addition, they noted improvements when combining LLM evaluation with linguistic feature-based approaches. The study demonstrated zero-shot LLM scoring is viable, but still needing methodological refinements to ensure consistency and construct validity.

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 (`https://chatgpt.com/`), and LLaMA-2 (`https://www.llama.com/llama2/`) in scoring written essays. Through various prompt-engineering tactics, they found that both models exhibited comparable performance, 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. Building on this, Pack et al. (2024) highlights while LLMs such as GPT models can achieve human-comparable scoring accuracy, performance varied across time and task type, raising questions about reliability and generalizability. Importantly, they highlight that while LLMs offer a significant advancement over earlier rule/feature-based AWE, systemic validity checks remain necessary to ensure LLMs capture more than superficial patterns in student writing. Additionally, Li and Liu (2024) also demonstrated that LLMs can match or exceed the performance of traditional AWE systems, while also providing more flexible and context-sensitive evaluations of student writing.

Alternatively, fine-tuning LLMs, even older or smaller models, has been proven more effective than directly prompting more advanced LLMs (Li and Ng, 2024). Wang and Gayed (2024) fine-tuned the GPT-3.5 model on a corpus of TOEFL argumentative essays with human benchmark scores and demonstrated that the fine-tuned models achieved higher scoring accuracy than zero-shot prompting of GPT-3.5 or GPT-4, while also being robust when scoring essays from unseen prompts. Gayed (2025) builds on this, by developing a freely accessible web-based application called AiAWE which uses an open-source LLM (Gemma-3-27B-IT) with Low-Rank Adaption (LoRA) adapters and rubric-guided system prompting to score argumentative essays. Through this system, the author was able to show both strong agreement with gold-standard ETS raters across standard metrics in addition to providing actionable deep-level rubric-aligned feedback to users. Another study by Cai et al. (2025) introduced the Rank-Then-Score (RTS) framework, which first ranks essays using a fine-tuned LLM, then scores them based on the rankings which outperformed traditional supervised fine-tuning techniques, particularly in Chinese datasets. Furthermore, Chu et al. (2024) proposed the Rationale-based Multiple Trait Scoring (RMTS) model, which 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 the rubric.

Despite the potential of LLMs to generate rich, formative feedback, the majority of recent AWE research has continued the traditional focus on scoring accuracy (Li and Ng, 2024). This persistent emphasis is likely because feedback evaluation is significantly more time-consuming. Unlike scores, which can be validated against quantitative benchmarks, the quality of feedback must be assessed

through qualitative, pedagogical judgment. As a result, the use of LLMs for feedback remains largely anecdotal rather than proven through rigorous research. The limited number of studies that do cover feedback generation, such as those exploring Chain-of-Thought prompting to produce comments alongside scores (Stahl et al., 2024) or generating trait-specific rationales (Chu et al., 2024), represent important but initial steps. As such, this research aims to develop and validate an AWE system designed to provide both reliable scores and pedagogically valuable, fine-grained feedback.

3 Methods

3.1 LLM choices and procedures

Scoring LLMs Both open-source and commercial models were evaluated for scoring to ensure broad applicability. For the open-source option, LLaMA-3.3-70B-Instruct [4] was selected, the top model in instruction following[5] at the time this research was conducted. For the commercial model, following suggestions in Wang and Gayed (2024), GPT-4o[6] was selected, as it was 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.

Surface-level LLMs For the surface-level feedback module based on direct prompting, it is divided into two stages: (1) corrective edits and (2) the generation of explanations for those edits. In a separate study (Labib et al., 2027), DeepSeek-V3-671B[7], GPT-4o and LLaMA-3.3-70B-Instruct were compared in their GEC performance through zero-shot prompting using the W&I+LOCNESS dataset from the BEA 2019 Shared Task (Bryant et al., 2019). Findings (see Appendix G) suggest that the three models performed similarly across three commonly used evaluation metrics in GEC, including ERRANT $F_{0.5}$ (Bryant et al., 2017), GLEU (Napoles et al., 2016), and PT-ERRANT (Gong et al., 2022). Drawing on these findings, GPT-4o was chosen for the corrective edits as it performed better overall compared to LLaMA 3.3, while the small performance margin did not justify the significant infrastructure costs required to host the 671B DeepSeek model. To avoid overcorrection, the prompt to GPT-4o specifically asked it not to introduce stylistic corrections (prompt in Appendix D).

To generate explanations for the corrective edits, qualitative evaluation of three models, Claude 3.7 Sonnet (https://www.anthropic.com/news/claude-3-7-sonnet), GPT-4o, and Deepseek-R1 (https://huggingface.co/deepseek-ai/DeepSeek-R1), found that the explanations from all three models were of a similarly high quality (see Appendix B for example explanations). However, despite its strong performance, Deepseek-R1 was eliminated from consideration due to its significant latency in reasoning, which would negatively impact the user experience in an interactive system. Between the remaining two models, GPT-4o was selected to streamline the technical architecture, as using the same model for both corrections and explanations avoids the overhead of managing a second API call within the surface-level feedback module.

Deep-level LLMs For the deep-level feedback module, supervised fine-tuning with the curated dataset of 90 essays was tested on LLaMA-3.3-70B and GPT-4o. For LLaMA, the model was fine-tuned using an entropy-based loss function over 20 epochs (see Appendix A). The fine-tuning parameters of GPT-4o were unknown as it is a closed commercial model. Alongside fine-tuning we experimented with direct prompting for deep level feedback. Inspired by Chu et al. (2024), this research first conducted a thorough thematic analysis of the teacher feedback in the training subset (resulting multiple traits are shown in Table 1). Based on the two primary categories of macro (paragraph and essay level) and micro (word, phrase and sentence level) feedback, the deep-level feedback was further divided into two pipelines and few-shot prompts were engineered to instruct Claude 3.7, GPT-4o and LLaMA 3.3 to deliver feedback focusing on these subcategories in both pipelines (prompts in Appendix E).

3.2 Validation scheme

3.2.1 Score validation

For the scoring module, the fine-tuned LLaMA-3.3 and GPT-4o models were tested using the test subset of 360 essays and the same prompt as the fine-tuning prompt. To evaluate scoring accuracy, three metrics were employed: Root Mean Square Error (RMSE), Quadratic Weighted Kappa (QWK), and

[4] (https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct)

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

[6] https://platform.openai.com/docs/models/gpt-4o

[7] https://github.com/deepseek-ai/DeepSeek-V3

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 1: Categories for deep-level feedback

percentage agreement following Wang and Gayed (2024), which is the proportion of essays of exact agreement (same score) and adjacent agreement (a difference of 0.5) with benchmark scores.

3.2.2 Feedback validation

For surface-level feedback, each original essay and its corresponding corrected version were aligned 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 for:

- **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, a preliminary check on data integrity and sufficiency of feedback instances was first performed on output from the models. The prompt in the macro feedback pipeline asked the LLMs to specifically 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, inter-rater reliability was calculated. For any cases where the two primary raters disagreed, a third expert rater was consulted to deliver the final judgment.

3.3 Dataset and subsets

As this study required a dataset of argumentative essays with benchmark scores and comprehensive feedback, a proprietary dataset of 480 TOEFL Independent Writing test-taker samples with official scores from ETS[9] was obtained. 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[10]. 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 is shown in Table 2.

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 2: Score distribution in the dataset

For the scoring module, 120 essays were selected through score-based stratified sampling across the two essay prompts as the fine-tuning subset. And the remaining 360 essays comprised the test subset.

For the feedback module, as no feedback was available in the dataset, this research curated its own datasets for comprehensive feedback. For surface-level feedback in the form of corrective

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

[9] Access to this dataset is restricted to researchers approved by ETS under a limited, non-exclusive, revocable, and non-transferable license.

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

edits to grammar and mechanics, previous research on GEC has shown that directly prompting LLMs generates satisfactory results (Davis et al., 2024; Labib et al., 2027). As such, supervised fine-tuning and the creation of corresponding datasets were not necessary. For deep-level feedback that requires higher-order thinking, 90 essays from the 480 essays were selected using score-based stratified sampling. Eight experienced university teachers who teach academic writing courses were recruited to provide feedback using Microsoft (MS) Word's comment feature (see Figure 1 for an example of the comments). Before annotation, the teachers received training on the scoring rubrics and were given nine benchmark essays scored from 1 to 5 and annotated by the researchers 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. Their feedback data was then processed to create the fine-tuning subset for deep-level feedback.

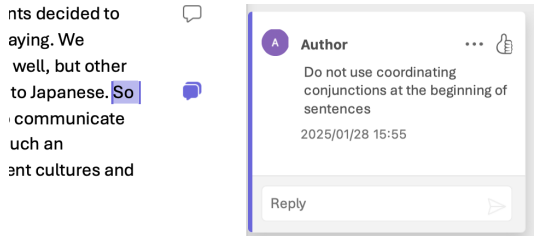


Figure 1: Example of a teacher comment in MS Word

For testing of both surface-level and deep-level feedback, due to the extensive human labor in evaluating such feedback, only 40 essays were selected (through stratified sampling) from the 390 non-annotated essays (dataset summary in Appendix F).

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), a prior study that used the same dataset.

Results show our two fine-tuned models improve on the baseline (Wang and Gayed, 2024), 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). 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 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 not deemed an error by raters.

4.3 Deep-level feedback module

4.3.1 Preliminary check

Preliminary check showed that neither fine-tuned models generated appropriate output data. The fine-tuned GPT-4o model produced incomplete output for every inference as it generated so many comments that the output token exceeded the context window of 8,000 tokens. Meanwhile, the fine-tuned LLaMA-3.3-70B model produced formatting errors, that prevented its output from being parsed as valid JSON. The most severe issues were structure errors, such as missing key-value delimiters and unclosed braces. As such, feedback from fine-tuned models were excluded from human evaluation.

For directly prompted models, preliminary check on feedback instances revealed that the micro feed-

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

back comments from LLaMA-3.3-70B and GPT-4o were insufficient. Of the 40 test essays, LLaMA-3.3-70B generated 336 comments and GPT-4o generated 368 comments, the averages of both were below 10. Claude 3.7, however, generated 630 comments. To verify that this was a difference in quality and not just verbosity, qualitative review of three randomly selected essays was conducted, which confirmed that the additional comments were indeed more fine-grained and pedagogically necessary. Given the intensive human labor required, human validation efforts were focused on the feedback from the most promising model, Claude 3.7.

4.3.2 Macro feedback

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: Macro comment evaluation by two raters

To assess inter-rater reliability, Gwet's $AC1$ coefficient (Gwet, 2008) was employed, 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 (Wongpakaran et al., 2013). 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. 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), under which the observed value indicates almost perfect agreement ($p < .001$).

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, 14 (6.97%). For the 14 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 combined with the previous one (N=8). In addition, some test-takers were unable to complete the task within the allotted 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).

4.3.3 Micro feedback

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 shown in Table 5, while that on the effectiveness of necessary comments mutually agreed by two raters is shown in Table 6.

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

Table 5: Necessity of micro feedback by two raters

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

Table 6: Effectiveness of judgment for micro feedback

Using Gwet's $AC1$ (Gwet, 2008), it is shown that the agreement on necessity was statistically

significant ($AC1$= 0.91; SE=0.01; 95% CI, [0.89, 0.94]; $p < .001$), as well as the agreement on effectiveness ($AC1$= 1.00; SE=0.00; 95% CI, [0.99, 1.00]; $p < .001$). After a third rater resolved the disagreements, the number of necessary comments was finalized as 600 (95.24%). 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 expands 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.

5 Discussion

As shown in Figure 2, the architecture was finalized by selecting LLMs tailored to each module: 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. To mimic human feedback mechanisms, an interactive UI was also developed (details in Appendix H).

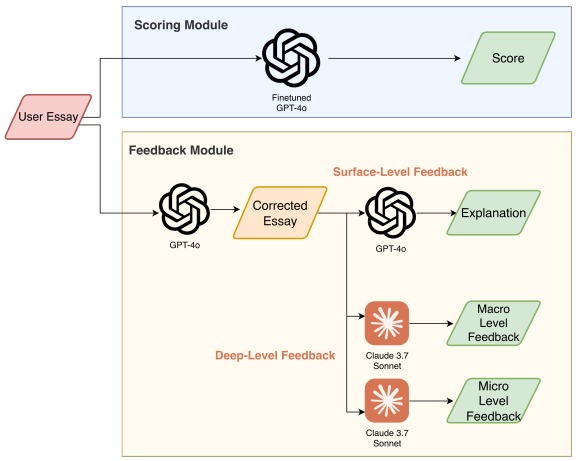


Figure 2: Final system architecture

The findings in this study are supported 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 fine-tuning, which is supported by the findings of Zeng et al. (2024). Particularly, as shown by the high precision rate of 96.14%, 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).

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 fine-tuned models. Here, fine-tuning did not yield satisfactory results, because the fine-tuned GPT model provided feedback to almost every word, phrase, and sentence, most of which were unnecessary. As the fine-tuning parameters were unknown, it can only be speculated that supervised fine-tuning with non-deterministic output is not well-suited for GPT-4o. For the fine-tuned LLaMA model, despite the data parsing issues, it was 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. Data formatting issue is likely an inherent issue with LLaMA, as previously highlighted by Yao et al. (2025). If the data formatting issue can be addressed, it may be possible to effectively use fine-tuned LLaMA for generating structured feedback.

Pedagogical implications The system can 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 help with 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.

6 Conclusion

By decomposing AWE into scoring, surface-level feedback, and deep-level feedback, *WrAFT* achieved strong performance in scoring on TOEFL writing dataset with satisfactory results on human validation of system-generated feedback. These findings provide a strong foundation for research in AWE systems that provide scoring and feedback.

Limitations

There are a few limitations in the design of this study. First, the system was developed 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 the validation scheme, a key limitation is that human evaluation only focused on precision without considering recall. For both surface-level and deep-level feedback, the necessity and effectiveness of feedback were assessed but it was not evaluated 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. But this research provides a necessary contribution for future studies to focus on measuring actual impact on student learning outcomes by including 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, 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. Second, 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.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI assistants for grammar and spelling checking. All suggestions were reviewed and edited by the authors, who take full responsibility for the content of the publication.

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, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

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.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

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*.

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.

Paul Deane. 2013. On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1):7–24.

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.

John Maurice Gayed. 2025. Aiawe (version 1.0) [web application]. Accessed October 22, 2025.

Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. 2022. Revisiting grammatical error correction evaluation and beyond. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6891–6902, Abu Dhabi, United Arab Emirates. 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, Italy. ELRA and ICCL.

Adnan Labib, Qiao Wang, Yixuan Huang, and Zheng Yuan. 2027. Multi-dimensional evaluation of llms for grammatical error correction. In *Artificial Intelligence in Education*, pages 536–545, Cham. Springer Nature Switzerland.

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.

Wenchao Li and Haitao Liu. 2024. Applying large language models for automated essay scoring for non-native japanese. *Humanities and Social Sciences Communications*, 11(1):1–15.

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*.

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):100050.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. Gleu without tuning. *arXiv preprint arXiv:1605.02592*.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

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. In Mark D. Shermis and Jill C. Burstein, editors, *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Mahwah, NJ.

Les Perelman. 2014. When "the state of the art" is counting words. *Assessing Writing*, 21:104–111.

Donald E Powers, Jill C Burstein, Martin Chodorow, Mary E Fowles, and Karen Kukich. 2002. Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2):103–134.

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 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 (EMNLP 2016)*, pages 1882–1891, Austin, Texas, USA. Association for Computational Linguistics.

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

Qiao Wang and John 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, et al. 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.

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.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California. Association for Computational Linguistics.

Zheng Yuan, Shiva Taslimipoor, Christopher Davis, and Christopher Bryant. 2021. Multi-class grammatical error detection for correction: A tale of two systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8722–8736, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 Model and Training Configuration

The model was developed by applying full-parameter fine-tuning to LLaMA 3.3 70B. It was trained to generate fine-grained feedback on student argumentative essays. The training parameters are presented in Table 7.

Parameter	**Value**
Base model	LLaMA 3.3 70B
Fine-tuning method	Full-parameter fine-tuning
Number of epochs	20
Batch size	32
Learning rate	3×10^{-6}
Optimizer	AdamW
Learning-rate scheduler	Cosine scheduler

Table 7: Training parameters

The tokenizer parameters are presented in Table 8.

Parameter	**Value**
Tokenizer class	`PreTrainedTokenizerFast`
Configured maximum length	8,192 tokens
Model input fields	`input_ids`, `attention_mask`
Padding side	Right
Beginning-of-sequence token	`<|begin_of_text|>`
End-of-sequence token	`<|eot_id|>`
Padding token	`<|finetune_right_pad_id|>`
Additional special tokens	`<|eot_id|>`, `<|eom_id|>`
Clean up tokenization spaces	True
Split special tokens	False

Table 8: Tokenizer parameters

B Explanations of surface-level corrective edits by GPT-4o, Claude 3.7 and DeepSeek-R1

ID	Original / Correction	Claude 3.7 Sonnet	GPT-4o	DeepSeek-R1
0	*serve → serves*	Subject–verb agreement error. "Academic writing" is singular and requires "serves".	Subject–verb agreement error. "Writing" is singular, so verb should be "serves".	Subject–verb agreement error. "Academic writing" is singular and requires "serves".
1	*corner → cornerstone*	Word choice improvement. "Cornerstone" conveys the foundation or essential component of scholarly communication.	"Corner" is semantically incorrect; "cornerstone" is the correct metaphor for something fundamental.	"Corner" is incomplete; "cornerstone" better conveys fundamental importance.
2	*enable → enables*	Subject–verb agreement. "It" is singular and requires "enables".	Subject–verb agreement. "It" is singular, so verb should be "enables".	Subject–verb agreement. "It" is singular and requires "enables".
3	*they're → their*	Incorrect homophone. "They're" (they are) should be replaced with possessive "their".	Incorrect homophone. "They're" is incorrect; should be possessive "their".	Wrong word choice. "They're" (they are) vs. "their" (possessive).
4	*accurate → accurately*	Incorrect part of speech. Adverb "accurately" modifies verb "convey".	Adverb needed. "Accurate" → "accurately".	Adverb required. "Accurate" should be "accurately" to modify verb.
5	*spends → spend*	Subject–verb agreement. Plural "Scholars" requires plural "spend".	Subject–verb agreement. "Scholars" is plural → "spend".	Subject–verb agreement. "Scholars" plural → "spend".
6	*hone → honing*	Verb form consistency. "Honing" maintains parallel structure with "learning".	Verb form error. Change to gerund "honing" for parallelism.	Gerund required after "years". "Honing" shows ongoing action.
7	*learn to balance technic → learning to balance technical*	Verb form consistency & spelling. "Learning" parallels "honing"; "technic" → "technical".	Grammar and clarity. "Learn to balance technic" → "learning to balance technical".	Parallel structure and word correction. "Learning" matches "honing"; "technical" corrects "technic".
8	*discipline → disciplines*	Pluralization. Remove comma; use plural "disciplines".	Number agreement. "Many" → plural "disciplines".	Plural required. "Many" → plural "disciplines".
9	*help → helps*	Subject–verb agreement. "Process" is singular → "helps".	Subject–verb agreement. "The peer review process" is singular → "helps".	Subject–verb agreement. "Process" is singular → "helps".
10	*distribute → distributed*	Verb form. Passive voice requires past participle "distributed".	Verb tense. Use past participle "distributed" in passive construction.	Passive voice requires past participle. "Is distributed" correct form.

Table 9: Explanations of corrective edits generated by Claude 3.7 Sonnet, GPT-4o, and DeepSeek-R1.

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 Deep-level prompts

E.1 Prompt for macro feedback

Instructions:

You are an English teacher in argumentative writing and your job is to give macro feedback to a student's argumentative essay structure. An essay should typically include three major sections: Introduction, Body and Conclusion.

The Introduction part should open with some background of the essay topic, followed by a thesis statement of the author and possibly end with a predictor of sub-points for later paragraphs.

The body part should contain detailed argumentation of sub-points of the thesis statement, probably divided into several paragraphs, with each paragraph having a clear topic sentence.

The conclusion part should efficient summaries the essay and preferably end with a forward looking statement.

You need to first identify the three sections, and then give feedback to each paragraph. Specifically, you need to identify the issues of essay structure targeting specific paragraph(s).

- In the introduction paragraph, there may be a lack of background information, thesis statement (main claim) or a predictor for following content, or the introduction does not effectively introduce the topic;
- In body paragraphs, there may be a lack of topic sentences and issues in argumentation. Specifically for argumentation, focus on ideas that are irrelevant to the thesis statement or the claim; repeated and thus redundant ideas, underdeveloped arguments where there is a lack of evidence and support, e.g, in the form of examples, or the support is poor or irrelevant; where there is a lack of warrant; where the author's stance is confusing; and where handling of counter-argument is poor, etc.
- In the conclusion paragraph, there may be undiscussed ideas, ideas that shouldn't be placed in the conclusion, such as further arguments or counter-arguments, and lack of effective summary of main points, etc.
- For all the three section, please consider whether proper paragraphing is there. Improper paragraphing may include putting content that should be in the body part, e.g., de-

velopment of main points, examples, counter-arguments, etc., with the introduction or conclusion part, and vice versa.

For each paragraph, write down feedback comments, including suggestions for solutions, possibly with an example of how to rewrite. Please make sure the number of comments your return is equal to the number of paragraphs and use the second person pronoun "you" to sound like you are talking to the student directly. Do not specifically divide your comments into "issue:" and "solution:" points; instead write it as a coherent piece of text.

Input and Output

You will be given the essay prompt, the essay text of the student in the user input. Your response should be in the following JSON format:

```
  {
"comments": [
  {
    "id": 0, // The id of the comment
    "Paragraph": , // The number of the
    specific paragraph
    "data": "...", // The teacher's feedback
    comment
  },
  ...
]
}
```

E.2 Prompt for micro feedback

You are an English teacher specializing in argumentative writing. Your task is to provide detailed micro feedback on a student's argumentative essay, focusing on grammar, clarity, formality, and coherence. You will analyze the essay and provide specific, actionable feedback for improvement.

```
Here is the essay prompt:
<essay_prompt>
{{essay_prompt}}
</essay_prompt>

And here is the student's essay:
<essay_text>
{{essay_text}}
</essay_text>
```

Please carefully read and analyze the essay, then provide feedback according to the following guidelines:

1. Focus Areas:

a) Grammar:

- Identify and correct grammatical and mechanical errors

- Address context-dependent grammar issues (e.g., appropriate tense, aspect)

b) Clarity:

- Highlight unclear, inaccurate, or redundant meanings

- Point out lengthy or confusing sentence structures

- Identify poor or unclear referents c) Formality:

- Note expressions requiring hedging

- Identify informal language use (e.g., coordinating conjunctions at sentence beginnings, subjective wording, informal interrogatives, colloquial language, second-person pronouns, contractions)

d) Coherence:

- Evaluate connections between sentences and ideas

- Assess transitions between paragraphs

- Check consistency in person perspective

2. Feedback Structure:

For each issue you identify, provide the following information:

- The type of issue (grammar, clarity, formality, or coherence)

- The specific paragraph number (starting from 0) and the micro text span (preferably words or phrases, avoiding clauses and sentences when possible)

- A clear explanation of the issue

- A suggested solution

3. Additional Guidelines:

- If similar issues occur multiple times, acknowledge this in your feedback (e.g., "As mentioned previously...")

- Address the student directly using the second-person pronoun "you"

- Write your feedback as coherent paragraphs rather than dividing into explicit "issue" and "solution" sections

- Ensure your comments are not repetitive

- Highlight the shortest text spans possible for feedback, unless a longer span is absolutely necessary

4. Output Format:

Your final output should be in JSON format, structured as follows:

```
{
  "comments": [
 {
   "id": 0, // The id of the comment
        "Paragraph": , // The number of the
        specific paragraph,
                  // starting from 0
   "start": 22, // The start character
   position of the
                // highlighted text in the
                paragraph
```

```
    "end": 36, // The end character position of
    the
               // highlighted text in the
               paragraph
    "highlighted_text": "...", // The
    highlighted text
                              // span to be
                              commented on
    "data": "...", // The teacher's feedback
    comment
                  // (explanation of the issue
                  and solution)
    "type": "..." // The type of issue, one of
    "grammar",
                  // "clarity", "formality",
                  "coherence"},
     ...// More comments...
  ]
}
```

Before providing your final JSON output, wrap your analysis inside <essay_analysis> tags in your thinking block. Follow these steps:

1. Grammar analysis:
- List specific grammatical errors found
- Note any context-dependent grammar issues
2. Clarity analysis:
- Identify unclear or redundant phrases
- List confusing sentence structures
- Note any poor referents
3. Formality analysis:
- List instances of informal language use
- Identify expressions that require hedging
4. Coherence analysis:
- Evaluate connections between sentences and paragraphs
- Note any inconsistencies in perspective
- Avoid commenting on general essay structure (general comments on introduction, body and conclusion paragraphs)

For each issue, consider multiple perspectives and potential solutions. This thorough analysis will ensure accurate and helpful feedback.

Begin your analysis now, followed by the JSON-formatted feedback. Your final output should consist only of the JSON-formatted feedback and should not duplicate or rehash any of the work you did in the thinking block.

F Dataset Splits

The dataset splits used for fine-tuning and testing are shown in Table 10

G GEC System performance

The GEC system performance, as stated in (Labib et al., 2027), is provided in Table 11

H Interactive UI

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 3). For the deep-level feedback tab, a comment-based interface design was adopted, 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 corresponding feedback comment appears on the right-side panel (see Figure 4).

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 10: Datasets used in fine-tuning and testing

Model	ERRANT	GLEU	PT-ERRANT
DeepSeek-V3-671B	0.4926	0.7677	0.5666
GPT-4o	0.4592	0.7420	0.5484
LLaMA-3.3-70B-Instruct	0.4826	0.7345	0.4973

Table 11: GEC Performance of prompting SOTA LLMs

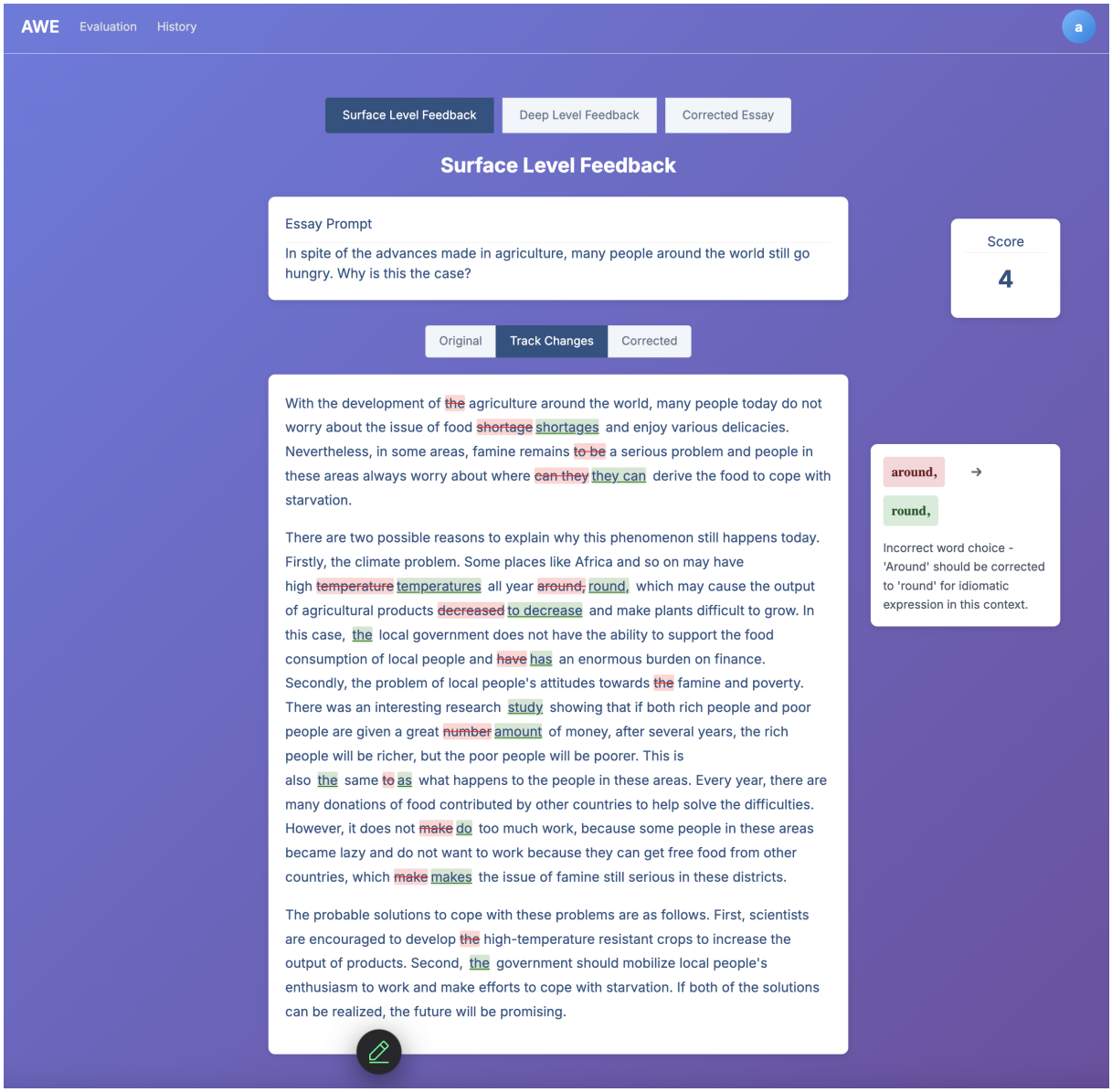


Figure 3: Surface-level feedback page - track changes

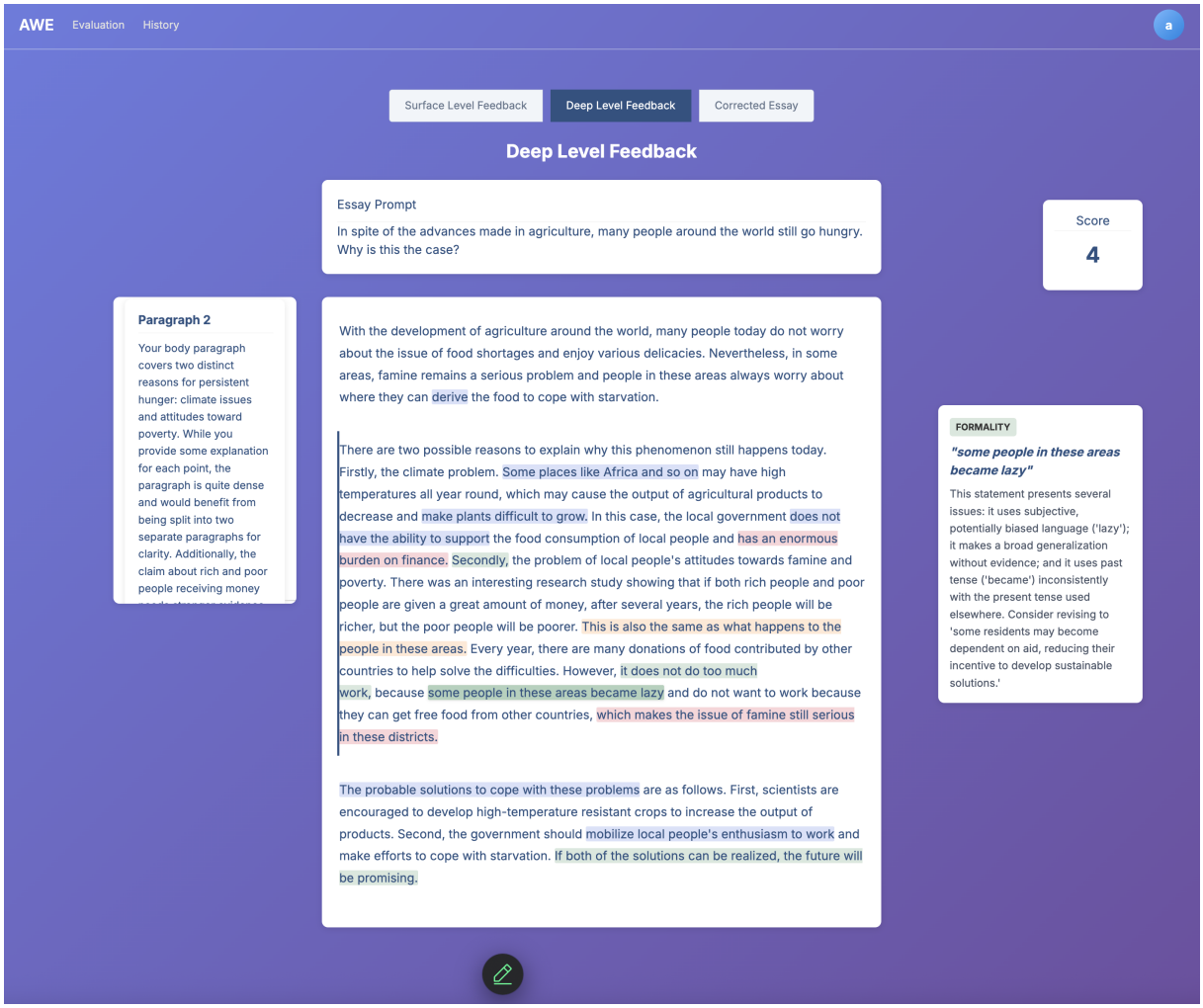


Figure 4: Deep-level feedback page