Operational Efficiency and Feasibility of Generative AI in English Writing Instruction:

A Study on Teacher-led Assessment Model

Kensaku Ishimaki [1]

Abstract

The increasing integration of Large Language Models (LLMs) offers potential approaches to addressing the persistent challenge of teacher workload in English writing instruction. This study explores a "Teacher-led AI Utilization Model," operating as a semi-autonomous workflow wherein the instructor assumes the role of a "System Architect." Utilizing a custom-built prototype system developed with Google Spreadsheet and Apps Script (GAS) alongside parallel prompt architectures (Claude-4-Sonnet and Gemini-Pro), the model was implemented in an advanced university English course (n=57). The research examines how structured system design and task re-engineering can align AI-generated outputs closely with pedagogical intent.

Operational observations indicated a degree of efficiency, enabling the finalization of comprehensive feedback for all submissions within three working days. However, comparative analysis highlighted a persistent need for human mediation. While the system achieved a 77.7% concordance rate at the individual evaluation item level, the rate of complete agreement per submission was limited to 10.5%. This discrepancy was primarily due to the AI's tendency toward "excessive rigor," necessitating instructor modifications for 89.5% of the raw scores.

These findings imply that while LLMs has potential to expand instructional capacity, the educator's role remains indispensable for pedagogical calibration and the final scrutiny of feedback. Ultimately, within the context of the proposed model, this study presents an example of how the role of the human instructor can evolve and deepen into a more advanced pedagogical dimension.

Keywords: Generative AI, English writing instruction, Automated writing assessment, Teacher-led model, System Architect.

[1] サイバー大学 IT 総合学部・准教授

1. Introduction

1.1 Background and Rationale

Since the emergence of ChatGPT at the end of 2022, the rapid spread of Generative AI, including Large Language Models (LLMs), has brought about a major paradigm shift in the field of education. English writing instruction is one of the areas experiencing the most significant impact.

In the field of automated writing assessment, Automated Essay Scoring (AES) existed as a conventional system. However, this was limited to evaluations focused on surface-level features based on the analysis of linguistic characteristics. The advent of LLMs, however, has demonstrated the potential to realize more flexible feedback, such as understanding context, interpreting logic, and even inferring and reflecting the learner's intent for individualized optimization.

In traditional writing instruction, the limit of the teacher's workload has always been a bottleneck in aiming to achieve both the immediacy of correction/grading and the qualitative comprehensiveness of feedback in large classes. When aiming to provide instruction optimized for individual students with varying basic academic abilities, comprehension, and proficiency levels, it inevitably resulted in immense operational burden and "Brain Fatigue" for teachers.

Within our institution, initiatives have been undertaken to address the workload bottleneck in large-scale, mandatory foundational courses, including the introduction of a feedback system that outsources primary corrections and evaluations (Fujisawa et al., 2024). While this outsourcing model has effectively supplemented human resources for massive cohorts, elective advanced courses, characterized by limited enrolment, do not necessarily require such external outsourcing. Furthermore, assignments in advanced courses tend to involve relatively more sophisticated integrated tasks and evaluation criteria. Consequently, given these distinct educational settings, a different approach from the aforementioned system was implemented.

Specifically, the integration of Generative AI was deemed a viable means to facilitate complex evaluations in advanced courses while reflecting the instructor's pedagogical intentions more directly. However, when corrections are made using general-purpose AI tools such as ChatGPT or Google Gemini with abstract prompts, risks such as "Over-correction"—which may not match the learner's proficiency level—or evaluations that diverge from the instructor's intent have been pointed out. Consequently, there is a concern that the intended learning outcomes may not be achieved.

Therefore, to effectively bridge the gap between educational objectives and AI-generated outputs, the present study introduces a prototype of a "teacher-led AI system". This experimental attempt aims to enable flexible operation tailored to the class size while mitigating the risks of generic AI use. In this approach, teachers take on the role of a mediator, applying their existing instructional insights to govern the behaviour of the technology. The 'Teacher-led AI Utilisation Model' is thus positioned as a viable methodology designed to ensure that AI serves as a high-fidelity extension of the teacher's intent.

1.2 Research Objectives

This study aims to position Generative AI as a core component of a system designed to augment instructional expertise. This approach seeks to enable higher-order instruction by simultaneously enhancing the quality and quantity of pedagogical feedback. Furthermore, the paper investigates the operational feasibility and efficiency of this model, focusing on the following three objectives:

- **First: Literature Review and Theoretical Framework**
  This study examines the transition from traditional Automated Essay Scoring (AES) to Generative AI-based evaluation, organizing the challenges inherent in this shift. This includes confirming the precision of AI in addressing both surface-level elements and deep context understanding, while establishing the validity of the "Hybrid Assessment Model" as a foundational best practice.
- **Second: Development of the System Architecture and the "Architect" Role**
  The research explores the practicality of an original system that utilizes dedicated prompts on a custom platform. This section focuses on the methodology of encoding pedagogical rubrics into system logic, illustrating how the instructor functions as a "System Architect" to ensure the integrity of the evaluation process.
- **Third: Empirical Analysis of Implementation and Teacher Mediation**
  Through a case study of an Advanced English course, the study evaluates the actual outcomes of the proposed workflow. By analysing the concordance between AI-generated assessments and instructor modifications, the analysis clarifies the necessity of human mediation and the model's impact on both instructional quality and operational efficiency.

1.3 Structure of the Paper

Based on the above objectives, this paper unfolds with the following structure.

- Chapter 2 reviews the literature on English writing instruction, automated scoring, and Generative AI. It outlines the transition from traditional AES to LLMs and

examines prior research regarding evaluation reliability.

- Chapter 3 reports on the implementation within an elective Advanced English course, detailing the development of a custom-built prototype scoring and correction system. It explains the architecture of a “semi-autonomous workflow” utilizing both Claude-4-Sonnet and Gemini-Pro, and assesses the operational efficiency achieved through AI-driven batch processing. Furthermore, it examines the specific evaluative characteristics of the system and the necessity of teacher mediation, based on an analysis of instructor modifications to AI-generated feedback drafts.
- Chapter 4 addresses the limitations of the current study and future prospects for the proposed model. By synthesizing the empirical results, it discusses how the model presents an example of instructors building upon their existing expertise to fulfil a more advanced pedagogical role in an AI-integrated environment.

2. Literature Review: The Evolution of Automated Writing Assessment

2.1 The Shift from Automated Essay Scoring (AES) to Large Language Models (LLMs)

Prior to the emergence of Generative AI, exemplified by ChatGPT in 2022, the field of automated English writing assessment was characterised by Automated Essay Scoring (AES) systems, such as ETS’s “e-rater®. ” AES systems scored essays by statistically analysing and extracting pre-defined “linguistic features,” such as lexical diversity, sentence length, and mechanical errors including spelling and grammar. While these attempts achieved a moderate level of accuracy, balancing cost reduction in large-scale exams with evaluation consistency, scholars have noted that they fell short of interpreting higher-order abilities such as the overall logical structure, rhetorical intent, or content creativity (Hannah et al., 2023).

The advent of Large Language Models (LLMs) in the 2020s has been a gamechanger for this field. Indeed, English writing correction and evaluation are arguably one of the domains most profoundly affected by Generative AI. A defining characteristic of LLMs is their capacity to “interpret” the text holistically and generate context-aware feedback. This has enabled “logical dialogue regarding the learner’s arguments” and “proposals for vocabulary and expressions aligned with context and intent”—tasks deemed difficult for traditional AES. Consequently, the focus of assessment is shifting from “formal correctness” to “communicative validity” (Hannah et al., 2023).

2.2 Evaluative Precision and Tendencies in Generative AI

Research indicates that the accuracy of Generative AI in correcting English text varies significantly by error type. First, regarding "surface-level errors" such as subject-verb agreement, number agreement, spelling, and basic word order, AI models demonstrate high identification accuracy, exceeding 90%. Conversely, accuracy declines notably for items requiring "deep context understanding," such as unnatural sentence structures, logical cohesion via conjunctions, and interpreting the author's intent (Alsaweed & Aljebreen, 2024). Comparative studies between GPT-4 and experienced teachers confirm that, at present, AI reliability and accuracy remain inferior to human performance (Pack et al., 2025).

Second, the appropriateness of "correction proposals" poses a challenge. Due to the inherent nature of their generation mechanisms, Generative AI models produce highly fluent text. Consequently, there is a marked tendency towards "over-correction," wherein the learner's original expressions are substantially altered, often deviating from the original intent (Fang et al., 2024). In writing pedagogy, the principle of "Minimal Correction"—respecting the learner's voice as much as possible—is considered educationally desirable. However, studies note that unless constrained by precise prompts or few-shot examples, AI tends to rewrite text into overly fluent prose that overrides the learner's intent (Loem et al., 2023).

Third, concerns persist regarding the quality of AI-generated "explanations and commentary." Even where correction proposals are immediate and linguistically sound, they may not necessarily be pedagogically appropriate for the learner. Research highlights challenges regarding psychological acceptability, noting that explanations are frequently perceived as "ambiguous" or "lacking empathy" (Chen et al., 2024).

2.3 Assessment Reliability and the Rationale for the 'Hybrid Model'

Concerns regarding the reliability of Generative AI assessment centre on the issue of "model and prompt dependency". Given that evaluation outcomes are susceptible to fluctuations driven by frequent AI model updates and the specific phrasing of user prompts, relying solely on a single AI judgment requires cautious consideration. (Mizumoto et al., 2024; Guo & Wang, 2023).

Consequently, leading English testing organisations maintain that a "Hybrid Assessment Model"—combining primary AI evaluation with final judgment by skilled human experts—remains best practice. For instance, ETS has ensured reliability through concurrent use with humans since the early development of AES (Monaghan, 2005). Similarly, Cambridge English's "Linguaskill" employs a mechanism wherein

human experts intervene when the reliability of the AI judgment is low (Cambridge, 2022). Although the specific mechanisms for human intervention differ, both represent hybrid assessment models that perform judgments utilising both AI and humans in separate stages.

In keeping with this framework, this study adopts a hybrid approach wherein AI undertakes surface-level checks and the initial drafting of feedback. This allows the educator to concentrate on higher-order instruction—such as moderating assessment severity and ensuring constructive feedback tone—and final evaluation judgment. This represents a pragmatic strategy to maximise educational validity while mitigating the inherent limitations of current AI assessment models (Automated Writing Evaluation Tools, 2023; Examining AI-Based Accuracy Assessment, 2024).

3. Practice of Generative AI Utilization in Advanced Proficiency Classes: Integration of System Design and Task Development

3.1 Overview of Practice: Semi-Autonomous Evaluation and Feedback Generation

3.1.1 Concept and Implementation of the 'Semi-Autonomous Workflow'

This chapter reports on the implementation of Generative AI in an Advanced English class (1A) during the Spring 2025 semester. A defining feature of this practice is the establishment of a 'semi-autonomous workflow' in which AI is positioned as an assistant that undertakes the primary operational workload. Within this framework, the system generates the initial drafts of the feedback content, while the instructor finalizes the output through a process of review and editing.

In this advanced course, students are required to demonstrate the ability to execute complex 'Integrated Tasks'. These tasks involve, for instance, composing an English email by applying structural conventions learned from instructional content to information synthesised from an English source video (e.g., tips on airport logistics and cost-saving). Attempting to provide comprehensive and detailed feedback (FB) on such assignments independently would impose a significant cognitive and physical burden ('Brain Fatigue') on the instructor, while inevitably resulting in prolonged turnaround times.

In this practice, the 'semi-autonomous workflow' refers to a system wherein the instructor functions as an 'Architect', designing a process that enables the AI to replicate the instructor's pedagogical judgments as faithfully as possible. Consequently, the system drives the operational workflow—from generating evaluations and corrections to shaping feedback comments—under the architect's design. By rigorously

encoding the instructor's assessment logic into the system, a systematic automated workflow was established. As a result, this system enabled the generation and finalisation of comprehensive feedback sheets for all 57 submissions within three actual working days—a feat unfeasible for a solitary instructor using traditional methods.

This initiative demonstrates a novel approach in which the instructor's conventional expertise in correction and evaluation is extended and deepened into the realm of AI supervision and pedagogical calibration. In this context, "architecting" the system represents a higher-order extension of traditional teaching skills, allowing for a concentrated focus on complex evaluative decisions and the final refinement of AI-generated outputs.

3.1.2 Ethical Considerations

In this study, strictly anonymized data was used for analysis. The processing of student submissions via Generative AI tools (Claude and Gemini) was conducted in accordance with institutional data handling guidelines, ensuring that no personally identifiable information was exposed to model training data.

Regarding transparency in the instructional process, the utilization of Generative AI within the assessment workflow was not explicitly stated to the students. This approach was adopted to establish the AI strictly as a backend administrative tool to support the instructor. Therefore, it was deemed ethically imperative that the instructor assume full and unmitigated authorship responsibility for all feedback content delivered.

While Generative AI was integrated into the assessment workflow, its role was strictly limited to the generation of primary drafts for correction and evaluation. Every final feedback comment and score underwent a comprehensive review and modification process by the instructor. This "Human-in-the-Loop" framework ensures that all outputs represent the instructor's professional judgment and pedagogical intent. Furthermore, the feedback examples presented in the Appendix are synthesized samples created to protect student privacy.

3.2 Design and Development of an Original Scoring and Correction System

3.2.1 System Architecture

For this implementation, a proprietary scoring and correction system was constructed using Google Spreadsheet (GSS) and Google Apps Script (GAS) as the platform. The design attempted to synchronise the pedagogical intent of the task with the behaviour of the AI models. The operational workflow executed during the Spring

2025 semester is outlined below (see Figure 1).

1. **Primary Processing: Automated Batch Execution via GSS and GAS**
   Following the conclusion of the submission period, all student answer sheets were retrieved from the school LMS. The response sections from these individual submissions were consolidated into a central GSS to facilitate the subsequent single batch operation. The system then executed independent prompts concurrently for each of the ten assessment criteria (e.g., Subject, Address, Greetings, Content Points) using Claude-4-Sonnet. This process generated evaluations, correction proposals, and explanatory commentary collectively for the entire cohort.
2. **Secondary Processing and Final Confirmation: Teacher-Mediated Synthesis and Refinement**
   The discrete evaluation components generated during primary processing were manually transferred by the teacher into the chat interface of a 'Custom Gem' (a dedicated AI assistant) on Google Gemini for each individual submission. Using this interface, the teacher synthesised the raw data into cohesive feedback comments, refining the tone and content before finalising and recording them in the feedback sheet (see figure 1).

Secondary processing was initially intended to be purely manual. However, trials in a colleague's course revealed that most of the tasks were formal and repetitive, indicating they were well-suited for automation via Generative AI. After initially using general AI, a specialized 'Custom Gem' was developed for Advanced 1A to enhance

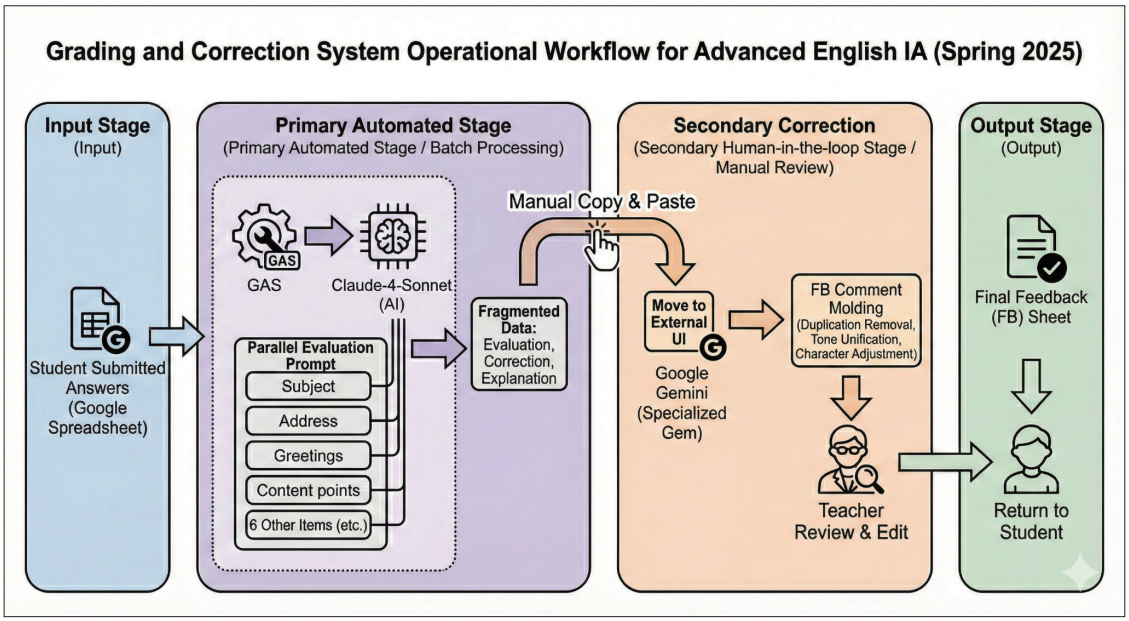

Figure 1: Operational Workflow of Scoring and Correction System in Advanced English IA (Spring 2025)

efficiency. These insights led to the decision to integrate the Gem-based processing into the batch workflow for the Autumn 2025 semester, leaving only the final review and editing as a manual stage for the instructor. While this refinement streamlines data handling, the Spring 2025 implementation serves as a proof-of-concept where manual review remains vital to pedagogical validity.

3.2.2 Parallel Prompt Architecture for 'Watertight' Evaluation

A central tenet of the system design was the assurance of logical robustness in evaluation, embodied by a 'Watertight' (leak-proof) design philosophy.

During the initial prototyping phase, attempts were made to process all evaluation criteria simultaneously via a single comprehensive prompt. However, this approach resulted in frequent 'leakage' due to 'attention dispersion', wherein the model failed to address multiple items consistently. To resolve this structural vulnerability, the implementation adopted a 'Parallel Evaluation Structure', in which the ten evaluation criteria are assessed independently. By executing dedicated prompts for each specific item, the architecture directs the LLM to concentrate its attention resources on discrete elements, thereby maximising both the comprehensiveness and accuracy of the evaluation (see Figure 2).

Specifically, evaluation criteria were deconstructed and explicitly articulated to eliminate interpretative ambiguity, aiming for a state where the AI could render consistent judgments. For instance, distinct judgment logic was embedded into the

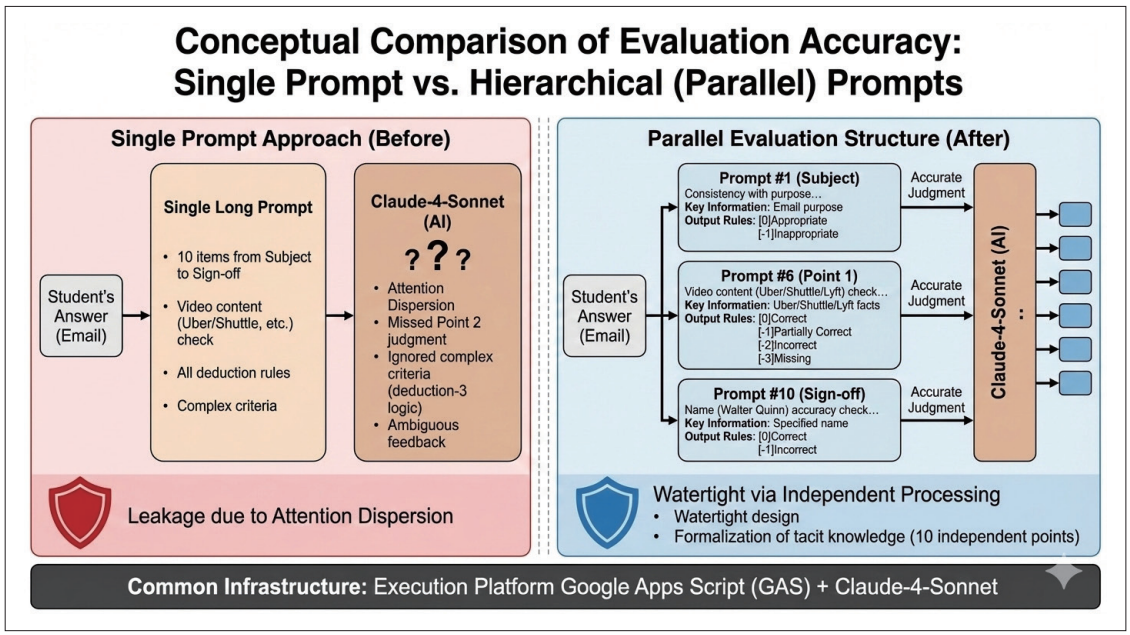

Figure 2: Conceptual Comparison of Evaluation Accuracy between Single Prompt and Hierarchical (Parallel) Prompts

prompts for each category, spanning from formal conventions such as 'Subject' and 'Address' to substantive 'Content Points' —specifically verifying the inclusion of accurate references to the video lesson. This configuration, akin to a 'checklist method', facilitated practical operation by significantly mitigating evaluation variability.

This process, wherein the instructor functions as a 'System Architect' to systematically encode pedagogical expertise into a suite of prompts, constitutes the core of the 'Watertight' design—a central theme of this study. This design philosophy aims to empower the AI to operate as a sophisticated evaluation engine, replicating the instructor's pedagogical intent to the best extent possible.

3.3 Task Redesign for AI Compatibility

3.3.1 Task Design as 'Requirements Definition for AI'

In this implementation, assignments were re-engineered as 'requirements definitions for AI'. This approach ensured a logical structure that facilitated AI processing while accurately reflecting educational intent.

- S**tipulating Persona and Scenario**: For the writing task, a highly specific context was established to reduce ambiguity. The student assumes the role of 'Walter Quinn, a new staff member at the travel agency Landhome', replying to a customer, Paige O'Brien, on behalf of a colleague, Cabe Dodd.
- **Codification of Structural Constraints**: Answers were strictly limited to '100–160 words', and specific formatting constraints, such as the 'prohibition of bullet points', were imposed to facilitate accurate automated scoring. Additionally, the instructions explicitly required the inclusion of specific references to the video content within the text.
- **Alignment with Prompt Architecture**: The study aimed to synchronise the pedagogical intent of the task with the behaviour of the AI model. This alignment was pursued by mirroring these task constraints within the scoring prompts, incorporating directives such as 'Respect the intent of the original text (Minimal Corrections)' and 'Provide explanations suitable for CEFR B1–B2 level students'.

3.3.2 Articulation of Assessment Rubrics and Integration of Materials

To ensure that AI judgments accurately reflect the instructor's pedagogical intent, efforts were made to explicitly articulate the underlying evaluation rubrics. Furthermore, the prompts, source materials, and instructional texts were configured to function as a cohesive 'tripartite' unit.

Specifically, 'Key words/information' were rigidly defined for each evaluation category (see Table 1). For instance, criteria verified the presence of formulaic

expressions in the introduction (e.g., "Thank you for contacting Landhome") and the inclusion of specific solutions within the content (e.g., "Uber, Lyft, or airport shuttles"). The central objective was to transfigure evaluation criteria that teachers traditionally assess 'intuitively' into concrete, AI-reproducible indicators—effectively converting 'tacit knowledge' into 'explicit knowledge'.

3.3.3 Establishing 'AI Resistance' through Integrated Task Design

To mitigate the risk of students bypassing learning objectives via general-purpose Generative AI, the structure of the 'Integrated Task' was fortified to necessitate the comprehension of external materials.

- **Mandatory Synthesis of External Sources**: Students were required to view specific video content on the university's e-learning platform, CU ENGLISH, and integrate distinct information retrieved from it (e.g., alternatives to public transport and cost-saving tips) into their responses.
- **Requirement for Personalised Input**: Beyond synthesising video content, instructions mandated the inclusion of 'original advice' tailored to a specific scenario—specifically, a parent travelling with a five-year-old child.
- **Implementation of Verification Logic**: Within the scoring prompts (Points 1 and 2), logic was implemented to strictly cross-reference student responses against the source material. This ensured the inclusion of specific solutions mentioned in the video (e.g., Uber, airport shuttles) while penalising the presence of inaccurate or 'hallucinated' information.

Through this strategic design, the task incorporated a level of 'AI resistance' — rendering it impossible to achieve high scores via simple, generic prompting—while

Table 1: Excerpt of Independent Evaluation Criteria and Design Rationale

Evaluation Item (Category)	Required Elements / Judgment Logic	Assessment Objective
Subject (Form)	Logic: Keywords: Travel Tips, NYC, Family Trip	Intent: Verification of conciseness and relevance regarding the email's purpose.
Point 1 (Content)	Logic: Required Lexis: Uber, Lyft, airport shuttles	Intent: Validation of accurate synthesis and comprehension of the source video.
Point 3 (Application)	Logic: Input Requirement: Original advice for traveling with a child	Intent: Assessment of original contribution beyond the source material.
Sign-off (Form)	Logic: Fixed Phrase: Best regards, Walter Quinn	Intent: Confirmation of adherence to the designated persona.

establishing an environment where the assessment AI could execute precise judgments as the instructor's effective 'alter ego'.

3.4 Analysis of Implementation Results

3.4.1 Comparative Analysis of AI-Generated Assessment and Instructor Evaluation

An analysis of the concordance between the AI's automated scoring and the instructor's final judgment revealed distinct tendencies in the system's evaluative performance, particularly when contrasting item-level accuracy with submission-level agreement.

First, when analysed at the "item-level," the system executed a total of 570 discrete evaluation instances (57 submissions × 10 independent criteria). Among these, the AI's judgments completely aligned with the instructor's final evaluations in 443 instances, resulting in an overall item-level concordance rate of 77.7% (see Table 2). This level of concordance suggests that the "watertight," parallel prompt architecture enables the AI to execute specific, compartmentalized assessment tasks with a certain degree of precision.

However, a stark contrast emerges when the data is examined at the "submission-level" (or essay-level). The rate of complete agreement—where all 10 criteria within a single submission matched the instructor's judgment perfectly—was observed in only 10.5% of cases (6 out of 57 submissions). Conversely, in the remaining 89.5% (51 submissions), the raw score (out of a maximum of 12) necessitated at least one modification by the instructor. Furthermore, even regarding the converted grade scale (out of 5), modifications were required for 75.4% (43 submissions) of the scripts.

The observed difference between the item-level concordance (77.7%) and the submission-level complete agreement (10.5%) highlights a characteristic of the current AI assessment model: while the AI demonstrated a certain degree of accuracy in individual verification tasks, cumulative discrepancies tend to arise when synthesizing multiple criteria into a holistic evaluation. This reaffirms the necessity of the instructor's final scrutiny.

Regarding the nature of these modifications, a clear tendency towards "excessive rigour" was confirmed. The AI applied assessment criteria with significantly greater strictness than the instructor, consistently calculating lower scores. Consequently, the overwhelming majority of instructor interventions involved upward adjustments to the grades. Where the AI mechanically applied deductions based on rigid criteria, the instructor intervened to interpret context and learner intent, thereby recalibrating the score.

Table 2: Concordance Rates between AI Assessment and Instructor Judgment by Criterion (Advanced IA, N=57)

Evaluation Criterion	Concordance Rate (Unmodified)	Upward	Downward	Operational Observations
1. Subject	98.25%	1	0	Binary verification of format yielded near-perfect precision.
2. Address	71.93%	15	1	Ambiguity in name entity recognition. Resolvable via prompt calibration.
3. Greetings	85.96%	7	1	—
4. Self Introduction	92.98%	4	0	Verification of mandatory information presence demonstrated high consistency.
5. Purpose	91.23%	5	0	—
6. Point 1 (Video Content 1)	56.14%	23	2	Increased divergence observed due to complexity of multi-level scoring (4-point scale).
7. Point 2 (Video Content 2)	63.16%	17	4	(Same as above)
8. Point 3 (Own Advice)	84.21%	5	4	—
9. Closing	77.19%	13	0	—
10. Sign-off & Name	56.14%	25	0	Discrepancies regarding persona adherence; necessitates alignment of prompt constraints with source material.
Total (N=570)	77.72% (442 items)	115	12	Overall performance confirms high individual item accuracy, contrasting with the lower complete agreement rate per essay.

However, a granular analysis reveals that this need for modification was not uniform across all criteria. While complex, multi-stage evaluation items often triggered the AI's excessive rigour, binary assessment items requiring simple verification—specifically 'Self Introduction' (Item 4) and 'Purpose' (Item 5)—demonstrated exceptionally high concordance rates.

3.4.2 Operational Efficiency and Mitigation of Instructor Burden

While a quantitative validation of the linguistic accuracy of AI-generated feedback falls outside the scope of this study, the operational reality—specifically the substantial reduction in working hours—warrants attention as a noteworthy outcome. Despite the necessity for score adjustments noted in Section 3.4.1, the deployment of this semi-autonomous workflow significantly mitigated the instructor's administrative and cognitive load, enabling the rapid completion of the feedback generation process—a

feat difficult to achieve through traditional solitary correction.

In terms of actual time allocation, the entire process was completed within approximately three working days. This duration comprised a Batch Processing Phase (approx. 4 hours) for the retrieval and pre-processing of 80 submissions across three courses, followed by an Individual Processing Phase (2 days) dedicated to the review and refinement of the 57 'Advanced English IA' submissions.

The primary driver of this efficiency was the AI's capacity to instantaneously generate detailed draft comments, which relieved the instructor from the burden of composing text de novo. This allowed for a redirection of attentional resources towards higher-order pedagogical tasks—specifically, moderating the AI's scoring severity and calibrating the tone of guidance to align with the course's target proficiency and the specific quality of the individual submission. The fact that the instructor could ratify the majority of these explanatory drafts with only minor refinements suggests the potential of the feedback to offer practical utility sufficient for higher education contexts.

3.4.3 Impact on Learner Perception and Perceived Utility

Although quantitative measurement of learning outcomes is outside the scope of this study, the course evaluation survey elicited favourable responses regarding the depth and precision of the feedback provided through this system. Although the number of specific qualitative comments was limited, students explicitly noted the quality of the guidance, with remarks such as "The feedback on the assignment was corrected very carefully" and "Mistakes were pointed out in detail."

For a representative example of the actual feedback architecture delivered to students, please refer to the Appendix (Note: To ensure privacy preservation, the data presented constitutes a synthesised example based on actual submissions and feedback content). It is inferred that the feedback of a volume and granularity previously unattainable for a solitary instructor directly correlated with the observed learner satisfaction.

3.5 Summary of Findings from the Advanced Course Implementation

The implementation within the Advanced English course demonstrated the operational feasibility of a semi-autonomous workflow. The key findings from the practice reported in sections 3.1 through 3.4 are summarized as follows:

- **Instructional Capacity and Efficiency**: By delegating the generation of primary feedback drafts to AI, the instructor finalized comprehensive feedback sheets for 57 submissions within three actual working days. This efficiency enabled the

delivery of detailed guidance at a volume and speed that would be difficult to sustain through traditional manual correction.

- **System Design and Task Integration**: The efficacy of the workflow was found to rely on the precision of prompt engineering to encode pedagogical expertise into "watertight" parallel prompt architectures. This structural measure suggested potential utility for compartmentalized assessment, with the system indicating a 77.7% concordance rate at the individual item level, demonstrating a certain degree of accuracy in executing discrete verification tasks.
- **The Necessity of Instructor Mediation**: While the system demonstrated the potential to significantly reduce instructor burden and shorten feedback creation time, the contrast between the item-level concordance and the 89.5% raw score modification rate per submission reaffirmed the vital importance of the human rater. This result indicates that cumulative discrepancies tend to arise when synthesizing multiple criteria into a holistic evaluation, underscoring the instructor's essential role in final scrutiny and the pedagogical calibration of the feedback tone.

In summary, the practice indicates that the instructor's role involves transmuting tacit pedagogical knowledge into explicit system logic. The system demonstrates the potential for drastically amplifying the instructional capacity, while it was also reaffirmed that the role of the human rater remains vital to this model.

4. Limitations and Future Prospects

4.1 Limitations of the Study

While the proposed model indicated a degree of operational efficiency, several limitations must be acknowledged:

- **Scope of Implementation**: The practical verification was limited to a specific course and a single instructor, which may constrain the generalisability of the findings to different instructional styles or settings.
- **Technical Dependency**: The evaluative performance remains highly dependent on the specific characteristics of the LLMs used (Claude-4-Sonnet and Gemini-Pro) and the precision of the current prompt engineering. Future model updates will likely require the recalibration of "watertight" architectures and prompts.
- **Absence of Long-term Learning Outcomes**: This study focused primarily on operational efficiency and assessment consistency. Further research is required to measure the long-term impact of AI-generated feedback on learner motivation and the actual improvement of writing proficiency.

4.2 Future Prospects and Concluding Remarks

The findings of this research suggest the possibilities of the following trajectories for English writing instruction in the AI era:

- **Expansion of Instructional Capacity**: While teacher workload has traditionally been a bottleneck in providing intensive writing instruction to large cohorts, the model presented in this study suggests that AI integration may have the potential to overcome these constraints. The model is expected to lead to enabling a single instructor to deliver immediate and individualized feedback at a volume that previously would have required a significantly greater investment of educational resources.
- **Course Design Premised on AI Feedback**: As an extension of this study, it is conceivable that individual instructors will be able to design their courses based on the premise that AI-generated feedback is available. For instance, within such a framework, an instructor could implement more frequent, iterative writing cycles, as learners would be supported by immediate and detailed guidance that encourages continuous refinement.
- **Elevation of the Instructor's Role**: By delegating a substantial portion of routine, repetitive, and administrative workloads to AI, instructors are able to concentrate their resources on higher-order responsibilities. These include sophisticated task design and high-level evaluative judgments that require an understanding of nuanced contexts. In this context, the role of the human educator is poised to evolve and deepen into a more advanced pedagogical dimension.
- **Challenges in Sharing Expertise Among Instructors**: However, if this system is to be used collaboratively by multiple instructors, a potential challenge arises from the isolation of specialized system design and operational know-how. Additionally, the possibility of discrepancies in the application and interpretation of assessment criteria, which is inevitable with multiple human raters, still persists. Therefore, the accumulation of shared knowledge among colleagues becomes essential.

The proliferation of Generative AI provides a significant opportunity to reflect on the indispensable role of the human instructor. The author concludes this paper with the hope that the "Teacher-led AI Integration Model" proposed herein will serve as a modest contribution to a new perspective on English education in an era of human-AI symbiosis.

Acknowledgements

The author extends sincere appreciation to Lecturer Jared Baierschmidt for his significant contributions to the development and practical operation of the system. Gratitude is also owed to Lecturer Fumi Wakui for her valuable advice during the writing of this paper. Finally, this study is presented as one of the academic outcomes based on the organizational management of courses within our institution. Any remaining errors or shortcomings are the sole responsibility of the author.

References

Aldosemani, T., Assalahi, H., Lhothali, A., & Albsisi, M. (2023). Automated writing evaluation in EFL contexts: A review of effectiveness, impact, and pedagogical implications. International Journal of Computer-Assisted Language Learning and Teaching, 13(1), 1–19. https://doi.org/10.4018/IJCALLT.329962

Alsaweed, W., & Aljebreen, S. (2024). Investigating the accuracy of ChatGPT as a writing error correction tool. International Journal of Computer-Assisted Language Learning and Teaching, 14(1), 1–18. https://doi.org/10.4018/IJCALLT.364847

Leading, L. L., Monaghan, W., & Bridgeman, B. (2005). E-rater as a Quality Control on Human Scores. *Connections*. https://www.kr.ets.org/Media/Research/pdf/RD_Connections2.pdf (accessed 8 January 2026)

Cambridge, C. E. M., & Dictionary, C. Technology-Enhanced Language Assessment. https://www.cambridgeassessment.org.uk/insights/technology-enhanced-language-assessment-innovative-approaches-for-better-learning/ (accessed 8 January 2026)

Chen, Q. (2024). Students' perceptions of AI-powered feedback in English writing: Benefits and challenges in higher education. International Journal of Changes in Education. Advance online publication. https://doi.org/10.47852/bonviewIJCE52025580

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. Education and Information Technologies, 29(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Fang, T., Yang, S., Lan, Y., Wong, D. F., Hu, J., Chao, L. S., & Zhang, Y. (2023). Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation. arXiv. https://doi.org/10.48550/arXiv.2304.01746

Fujisawa, K., Shirasu, Y., Sato, K., Kawachi, K., & Inoue, M. (2024). e-learning による英文ライティング指導のための結果フィードバックシステム導入 [Implementation of a result feedback system for English writing instruction via e-learning]. サイバー大学研究紀要 [Cyber University Research Bulletin], 1, 9–16. https://www.cyber-u.ac.jp/about/pdf/research_bulletin/001/CU_RB001_02.pdf

Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. Language Assessment Quarterly, 20(4–5), 399–420. https://doi.org/10.1080/15434303.2023.2288253

Lee, O. (2024). Examining AI-based accuracy assessment in L2 learners' writing. Journal of Pan-Pacific Association of Applied Linguistics, 28(2), 39–55. https://eric.ed.gov/?q=source%3A%22Journal+of+Pan-Pacific+Association+of+Applied+Linguistics%22&id=EJ1457668

Loem, M., Kaneko, M., Takase, S., & Okazaki, N. (2023). Exploring effectiveness of GPT-3 in

grammatical error correction: A study on performance and controllability in prompt-based methods. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 205–219). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2305.18156

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics*, *3*(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Pack, A., Hartshorn, K. J., Escalante, J., & Gillette, N. (2025). How well can GenAI (GPT-4) provide written corrective feedback on English-language learners' writing? International Journal of English for Academic Purposes: Research and Practice, 5(1). https://doi.org/10.3828/ijeap. 2025.2

Appendix

- Note: The following feedback sheet is a synthesized sample for illustrative purposes and does not contain actual student data.

2025SS Advanced English 1–A Writing Task

FB_code	25SS_AE1A_Sample	Raw Score	9
Student ID	Sample	成績評価点数	3

全体コメント

課題の基準を概ね満たしています。減点となったところをよく見直して復習しましょう。

Your Answer

Subject	About travel inquiry
Dear Ms. Paige, I hope this message find you well. My name is Walter Quinn, and I work as support desk at Landhome. I am writing to you behalf on my colleague Cabe Dodd who received your call yesterday. Regarding your family trip to New York City, I have some advoices for you. First, if you want to avoid public transport, I recommend you should choice Uber or a taxi to the airport. To save money during your travel, you should bring a refillable water bottle. Also, because you are traveling with a child, I recommend to bring some toys or books to keep him seated. I hope these suggestions are helpful. Please feel free to contact us if you have any questions. Best regards, Walter Quinn	

解説

#1 原文: I hope this message find you well.
添削後: I hope this message finds you well.
解説: 主語が「this message」（三人称単数）なので、動詞に三単現のsを付けると、文法的に正しくなります。

#2 原文: I am writing to you behalf on my colleague Cabe Dodd who received your call yesterday.
添削後: I am writing to you on behalf of my colleague Cabe Dodd who received your call yesterday.
解説: 「〜を代表して」という熟語は "on behalf of" です。正しい前置詞と語順をセットで覚えると、よりスムーズに書けるようになります。

#3 原文: I have some advoices for you.
添削後: I have some advice for you.
解説: adviceは不可算名詞（数えられない名詞）なので、常に単数形で使うように意識しましょう。

#4 原文: I recommend you should choice Uber or a taxi to the airport.
添削後: I recommend you choose Uber or a taxi to the airport.
解説: recommendの後は「主語＋動詞の原形」とするのが一般的です。名詞の choice ではなく、動詞の choose を使うと正確に伝わります。

#5 原文: I recommend to bring some toys or books to keep him seated.
添削後: I recommend bringing some toys or books to keep him seated.
解説: recommendの直後に動詞を置く場合は、動名詞（-ing）の形にしましょう。

評価の解説
[-1] 解説：宛名は、相手の姓を用いた "Dear Ms. O'Brien," としましょう。ビジネスメールでは姓（last name）を使用するのが標準的なマナーです。
[-1] 解説：課題の要件に正確に沿うため、本文では指定された選択肢（Uber、Lyft、またはairport shuttle）のみを用いるようにしましょう。
[-1] 解説：節約方法について、空港での飲食費の高さや荷物の預け入れ手数料を避ける点も具体的に説明しましょう。必要な情報を網羅することで、より説得力のある回答になります。

添削後の解答例（修正部分は [ ] で表示されます）

Dear Ms. Paige,

I hope this message [finds] you well. My name is Walter Quinn, and I work as support desk at Landhome. I am writing to you [on behalf of] my colleague Cabe Dodd who received your call yesterday. Regarding your family trip to New York City, I have some [advice] for you.

First, if you want to avoid public transport, I recommend you [choose] Uber or a taxi to the airport. To save money during your travel, you should bring a refillable water bottle. Also, because you are traveling with a child, I recommend [bringing] some toys or books to keep him seated.

I hope these suggestions are helpful. Please feel free to contact us if you have any questions.

Best regards,

Walter Quinn

英語ライティング指導における生成 AI の運用効率と実現可能性：

教員主導型評価モデルに関する研究

石　巻　賢　作

概　要

大規模言語モデル（LLM）の普及は英語ライティング指導に多大な影響を与えつつあり、添削評価のための教員の業務負担という長年のボトルネックが解消される可能性を示している。本研究では、教員が「システム・アーキテクト（設計者）」として機能する「教員主導型 AI 活用モデル」を提案し、これに基づく「半自動型ワークフロー」の運用可能性を検証した。Google スプレッドシートと Apps Script（GAS）、および並列プロンプト・アーキテクチャ（Claude-4-Sonnet および Gemini-Pro）を用いた独自のシステムを構築し、大学の上級英語コース（n=57）にて実践を行った。本稿では、特にシステム設計と課題の再構築を統合的に行ったことで、AI の出力をいかに教育的意図に合致させうるかを検証している。

実証結果として、全答案に対する包括的なフィードバックを実働 3 日以内に完了するなど、一定の運用効率が示された。一方で、AI による自動評価と教員の最終判断を比較した結果、人間による介在の継続的な必要性も明らかになった。評価項目単位での評価一致率は 77.7% であったのに対し、答案単位での完全一致率は 10.5% にとどまった。この差異は主に AI の「過剰な厳密さ（excessive rigor）」に起因しており、結果として全体の 89.5% の答案において教員による素点の修正が発生した。

これらの知見は、LLM が教員の指導キャパシティを拡張しうる一方で、教育的な配慮による調整やフィードバックの最終精査において、人間の教員の役割が依然として不可欠であることを示している。結論として本稿は、提案する AI 活用モデルの環境下において、人間の教員の役割がより高度な教育的次元へと発展・深化する具体的な一例を提示するものである。

キーワード：生成 AI、英語ライティング指導、自動添削評価、教員主導型モデル、
システム・アーキテクト