Indonesian Journal of English Language Teaching and Applied Linguistics

Vol. 11(1), May 2026
www.ijeltal.org
e-ISSN: 2527-8746; p-ISSN: 2527-6492

Gemini API-Based Automated English Paragraph Scoring Aligned with High School Thai Curriculum Writing Indicators

Phisit Deeboonmee Na Chumphae

Mahasarakham University, Thailand. email: phisitdeeboonmee@gmail.com

Received 6 February 2025 | Received in revised form 18 March 2026 | Accepted 20 March 2026

ARTICLE INFO

Keywords:
Automated Scoring, Gemini API, Paragraph Writing, Writing Assessment

DOI:
http://dx.doi.org/10.21093/ijeltal.v11i1.2553

ABSTRACT

This research aims to develop and evaluate an automated English paragraph scoring system using the Gemini API aligned with the Thai upper secondary curriculum's writing indicators, to address issues related to teacher workload and delayed feedback in writing assessments. The system integrates the Gemini 2.5 Pro API with a prompt-engineering framework designed to simulate expert EFL assessors. This research employs a sequential mixed-methods research approach. For the quantitative component, 160 upper secondary EFL students in Thailand were sampled from their written assignments, consisting of three expository paragraph assignments aligned with the Thai core curriculum. Cluster sampling was used to select participants. The students' writings were assessed using a validated analytical evaluation criterion comprising four aspects. The essays were independently scored by three evaluators, and the results were compared to automated scores generated by a Gemini-based system. Reliability between human evaluators was first checked using the Intraclass Correlation Coefficient (ICC), and the agreement between human and AI scores was measured using the Quadratic Weighted Kappa (QWK). The results showed a high level of agreement between the Gemini-generated scores and the human evaluators (QWK = 0.82), indicating that the system can approximate human judgment in evaluating English as a Foreign Language writing. Qualitative analysis of the AI-generated feedback further revealed that the system could provide diagnostic recommendations related to grammar, vocabulary, and sentence structure. These findings suggest that the system can support teachers in reducing grading workload while providing timely, criteria-based feedback to enhance students' writing development.

How to cite:
Na Chumphae, P. D. (2026). Gemini API-Based Automated English Paragraph Scoring Aligned with High School Thai Curriculum Writing Indicators. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 11(1), 47-68

1. Introduction

English writing has become a more important skill for communication and academic engagement in the 21st century. Through written language, people can share information, present ideas, and communicate diverse perspectives to readers in various contexts. However, for language learners, writing is often viewed as one of the most difficult productive skills. It requires learners to integrate linguistic knowledge with cognitive processes to create, organize, and refine ideas so they can be clearly communicated to a specific target audience. (Hyland & Richards, 2014; Langan & Albright, 2020).

In a cognitive perspective, writing is a complex problem-solving activity. Writers must take several processes simultaneously, such as planning ideas, choosing appropriate language, and anticipating how the reader will interpret the text. In addition, social and cultural perspectives emphasize that writing also functions as a social activity. Practical communication requires writers to use their language choices to different audiences and communication contexts in order to convey meaning successfully (Teng et al., 2022).

In Thailand, the importance of English language skills has increased significantly over the past few decades. That is a result of globalization and the country's participation in the ASEAN Economic Community (Jarunthawatchai & Baker, 2024). In response to these changes, the Thai Ministry of Education has placed greater emphasis on communication skills in the core curriculum of basic education (Ministry of Education, 2008). According to the foreign language learning standards, upper secondary students are expected to use English to communicate information, ideas, and opinions clearly and appropriately. The curriculum, therefore, aims to develop learners' ability to exchange information while demonstrating awareness of linguistic and cultural appropriateness in communication. In addition, it highlights the importance of responsible communication as a means of building constructive relationships within society.

Recently, Thailand's education policy has used principles from the Common European Reference Framework for Languages (CEFR). This policy emphasizes communicative skills more than grammar instruction. However, the use of these policies has been inconsistent. Vimala (2025) notes that inequality between urban and rural schools, such as a shortage of qualified teachers, and the lingering influence of traditional linguistic norms, may limit the effectiveness of the reforms. As a result, many schools continue to prioritize standardized test results over the development of student communicative abilities, creating a gap between national policy expectations and classroom practice.

Assessing writing ability also presents several challenges in the real-world classroom context. These challenges can be categorized into the psychological dimensions of measurement, operational dimensions, and pedagogical dimensions. From a psychological perspective, Hyland & Richards (2014) note that the complexity of writing and the inherent subjectivity of human grading can impact both the accuracy and reliability of writing assessments. Differences in test conditions, as well as the linguistic and cultural backgrounds of the assessors, may lead to variability in scoring that does not always accurately reflect students' writing abilities.

Operational challenges further complicate the assessment process. Traditional manual grading requires considerable time and effort. That often delays the provision of the feedback

students need to improve their writing. Tongsilp et al. (2024) and Huang and Wilson (2021) note that bias and inconsistencies among graders can exacerbate these problems. In particular, lengthy grading sessions can lead to mental fatigue, which may unintentionally alter grading standards and undermine grading fairness.

From a teaching perspective, giving feedback on writing is important for supporting learning. Harmer (2007, 2015) shows that correction alone is insufficient for developing writing skills. It can discourage learners without constructive feedback. Therefore, effective assessment requires feedback that students can understand and apply. In this way, students can improve their work and develop stronger writing abilities.

To decrease these challenges, researchers have explored Automated Writing Evaluation (AWE) systems supported by Large Language Models (LLMs). Previous studies show that these technologies can help teachers assess student writing more efficiently while providing immediate feedback to learners (Ajabshir & Ebadi, 2023; Wei et al., 2023). Among the emerging tools in this field, the Google Gemini API is an advanced AI model capable of accurate natural language processing and analysis. Previous research indicated that one of Gemini's key pros is its ability to interpret linguistic context and generate feedback similar to human responses (Mizumoto & Eguchi, 2023). This capability allows the AI system to give constructive feedback. That can encourage students to revise and improve their writing. As Imran & Almusharraf (2024), such feedback can foster self-directed learning. That allows students to experiment with language and refine their work through repeated revisions. In addition, many studies have shown that AI-powered writing tools may support improved writing consistency and overall learning outcomes compared to traditional instruction alone (Aliakbari et al., 2025).

Despite these promising developments, empirical studies examining AI-driven writing assessment systems in local educational contexts remain limited. Specifically, few studies have examined how well such systems align with Thailand's national curriculum standards. To address this gap, this study developed and evaluated an automated grading system powered by the Gemini API, specifically designed for assessing expository paragraph writing among Thai high school students. The study focused on writing at the paragraph level and examined how well the system could function in alignment with Thailand's basic education curriculum.

By exploring the creative applications of AI in this context, this study contributes to both theoretical and practical discussions in the field of writing assessment. From a theoretical perspective, the study provides empirical evidence of the reliability of AI-driven grading in local educational environments. From a practical perspective, the study proposes a scalable approach that could reduce the workload of grading teachers while providing timely and individualized feedback to students. Accordingly, this research addresses the following research questions:

RQ1: How can a Gemini API-based automated scoring system be developed to assess Thai EFL writing skills?

RQ2: To what extent is the scoring consistency of the system comparable to that of expert human raters?

2. Literature Review

2.1 Thai Basic Education Core Curriculum for English Productive Skills

In recent years, the Ministry of Education in Thailand has increasingly emphasized Communicative Language Teaching (CLT). Teachers are encouraged to use this approach to enhance students' effective language skills. This shift reflects broader efforts to align English language education in Thailand with the communicative principles of the Common European Framework of Reference for Languages (CEFR). As part of this reform, several initiatives have been introduced to promote more interactive and student-centered classroom learning. One example is the Boot Camp program, which was designed to shift English instruction toward communicative learning activities that encourage active language use among students. These policy directions are reflected in Standards F1.2 and F1.3 of the national curriculum. (Jarunthawatchai & Baker, 2024).

Ministry of Education (2008) created the learning objectives in the Core Curriculum within an important framework. In foreign language learning focuses on developing communication skills and fostering a positive attitude. Importantly, the use of English in an international context. To achieve these goals, the curriculum is divided into four interconnected sections, providing a guideline for English language instruction at different educational levels.

(1) Language for Communication focuses on developing the four basic language skills: listening, speaking, reading, and writing. These skills allow learners to exchange information, express opinions, and communicate ideas effectively.
(2) Language and Culture emphasizes cultural awareness. It encourages students to understand linguistic and cultural differences and to compare global practices with Thai cultural contexts.
(3) Language and Relationship with Other Learning Areas promotes the integration of language learning with other academic subjects. Through this approach, students can use English as a tool for accessing knowledge and expanding their academic perspectives.
(4) Language and Relationship with Community and the World highlights the practical use of language skills beyond the classroom. In this strand, learners are expected to use English in real-life situations such as further education, professional development, and international communication.

Table 1: Indicators and Core Learning Content of the Foreign Language Learning Area, Standard F1.2 Possessing language communication skills for the exchange of data and information, expression of feelings and opinions efficiently

Indicators	Core Learning Content
Converse and write to exchange data about oneself, various matters of near self, experiences, situations, news/incidents, and issues of interest to society, and communicate continuously and appropriately.	Language used for interpersonal communication, such as greeting, leave-taking, thanking, apologizing, complimenting, polite interruption, persuasion, exchange of data about oneself, matters of near self, various situations in daily life; conversation/writing data about oneself and people near oneself, experiences, various situations, and news/incidents.
Choose and use requests, give instructions, clarifications, and explanations fluently.	Requests, instructions, clarifications, and explanations with complex steps.

Speak and write to express needs; offer, accept, and refuse to give help in simulated or real situations appropriately.	Language used for expressing needs, offering and giving help, accepting and refusing help in various situations.
Speak and write to ask for and give data, describe, explain, compare, and express opinions about matters/issues/news/incidents that are listened to and read appropriately.	Vocabulary, idioms, sentences, and text used for asking for and giving data, describing, explaining, comparing, and expressing opinions about issues/news/incidents listened to and read.
Speak and write to describe their own feelings and express opinions about various matters, activities, experiences, and news/incidents with proper reasoning.	Language used for expressing feelings, opinions, and giving supporting reasons, such as liking, disliking, being glad, sorry, happy, sad, hungry, taste, beautiful, ugly, noisy, good, bad, regarding news, incidents, and situations in daily life.

Standard F1.2 indicators for upper secondary students, covering skills in exchanging information, expressing needs, and analyzing news with reasoning. It emphasizes using appropriate vocabulary and complex structures to communicate fluently in both simulated and real-life contexts. These standards serve as the framework for assessing students' proficiency in effective interpersonal communication.

Table 2: Indicators and Core Learning Content of the Foreign Language Learning Area, Standard F1.3 Presenting data, concepts, and opinions on various topics through speaking and writing

Indicators	Core Learning Content
Speak and write to present data regarding oneself, experiences, news/incidents, matters, and various issues of interest to society.	Presentation of data regarding oneself, experiences, news/incidents, matters, and issues of interest to society, such as journeys, dining, playing sports/music, watching movies, listening to music, raising pets, reading books, tourism, education, social conditions, and the economy.
Speak and write to summarize the main idea/theme identified from the analysis of matters, activities, news, incidents, and situations of interest.	Summarizing main ideas/themes; analysis of matters, activities, news, incidents, and situations of interest.
Speak and write to express opinions about activities, experiences, and incidents at local, global, and social levels, as well as provide justifications and examples.	Expression of opinions, providing justifications and examples regarding activities, experiences, and incidents at local, social, and global levels.

Standard F1.3 indicators for upper secondary students, covering skills in presenting data, summarizing main concepts, and expressing opinions with reasoning. It emphasizes the ability to analyze news, incidents, and social issues to communicate thoughts fluently with supporting justifications and examples. These standards serve as the framework for assessing students' proficiency in informative and argumentative presentation through speaking and writing. Consequently, this study specifically adopts standard F1.2 and standard F1.3 as the primary framework for developing the automated paragraph scoring system.

2.2 Automated Essay Scoring

Automated Essay Scoring (AES) refers to a computer system that automatically evaluates and scores written responses (Firoozi et al., 2023). In the past, AES systems were often criticized for their limited ability to assess writing quality. Most systems relied on surface-

level linguistic features such as word frequency, sentence length, and basic grammatical patterns. So, these systems often failed to catch deeper aspects of writing. These include text organization, semantic relationships, and overall consistency. Recent advances in natural language processing have importantly expanded the capabilities of automated scoring systems. In particular, the emergence of Large Language Models (LLMs) has enabled AES tools to analyze language in more complex ways. These models can identify writing such as complex linguistic relationships, interpret meaning in context, and assess the consistency of written text more efficiently than earlier feature-based systems. So, the LLM-driven approach has become a key for developing more accurate and reliable tools for evaluating student writing (Atkinson & Palma, 2025). In addition to generating numerical scores, automated writing technology can also provide constructive feedback. Shermis & Burstein (2013) describe Automated Essay Evaluation (AEE) as a framework to support writing development by integrating scoring with natural language processing techniques. In this approach, automation can provide diagnostic feedback that helps students identify linguistic problems. Then students can revise their work before receiving a final evaluation from the teacher.

These systems have also evolved over time. Earlier AES models often relied on traditional machine learning methods that depended on predefined linguistic features. However, newer systems rely on deep learning architectures that allow models to interpret contextual meaning, rhetorical structure, and discourse congruence in a manner similar to human language processing (Mizumoto & Eguchi, 2023).

These developments have important implications for writing assessment in large educational contexts. AI-assisted systems can make writing evaluation more scalable and efficient, particularly in EFL classrooms where teachers often face large numbers of student essays. Nevertheless, researchers caution that automated systems should not operate entirely without human oversight. Kim et al. (2025) argue that careful supervision is necessary to maintain scoring accuracy and pedagogical validity. In addition, the quality of AI-generated feedback is strongly influenced by prompt design. As Ouyang & Jiao (2021) explain, well-structured prompts can help align a model's internal reasoning processes with specific educational rubrics and learning objectives.

2.3 Writing Assessment

Weigle (2002) notes that the selection of an appropriate rating scale is a central element in writing assessment design. One key issue is that writing should be evaluated using a single overall score or by considering multiple aspects of the text. In academic writing, scoring scales are divided into three types: core attribute scales, holistic scales, and analytical scales. These scales differ primarily in two dimensions: the specificity of the work and the scoring method. Work specificity refers to the degree to which the scale is tailored to a specific type of writing, while the scoring method concerns whether the evaluation uses a single overall score or multiple scores for different aspects of the writing. Within this framework, core attribute scales are typically designed for specific works, while holistic and analytical scales are more broadly applicable to a wider variety of writing contexts.

The scoring method has important implications for the accuracy and usefulness of written tasks. Hyland & Richards (2014) state that holistic scoring is effective and practical for broad assessments, but generally provides limited diagnostic information for learners. In contrast,

analytical scoring assesses multiple aspects of a written work separately. That allows for the identification of specific strengths and weaknesses in the student's work. While this method generally requires more time and effort from the assessor, it often provides more reliable information for instructional feedback. Tesfay (2017) describes the importance of developing subject-specific rubrics, stating that effective assessment tools should reflect the specific objectives and standards of the educational context in which they are used.

Recent advances in artificial intelligence (AI) have expanded the possibilities for evaluating writing. In addition to focusing solely on final scores, AI-powered systems are increasingly integrating writing analysis that examines various dimensions of student learning outcomes. Techniques such as regression prediction, writing quality classification, and comparative ranking can be applied to analyze different aspects of student writing. These approaches allow automated systems to provide detailed feedback on attributes such as logical flow, stylistic choice, and sentence fluidity. This type of feedback can help shift the focus of writing evaluation from a single grade to a process of continuous improvement, potentially boosting students' motivation and confidence in their writing abilities (Swiecki et al., 2022).

Based on these components, this research uses analytical scoring as the primary assessment framework. Thailand's core curriculum places significant emphasis on assessing students' specific communication skills in written work. This is especially important for the relevance of content and clarity of expression. Therefore, analytical scoring is a suitable framework for adapting automated assessments to align with the expectations of these curricula. With this framework, the Gemini-based system can transcend conventional scoring and generate criteria-based feedback that directly reflects the learning objectives of Thai EFL students. These AI-based approaches extend traditional analytic scoring by enabling multidimensional evaluation of writing performance.

2.4 Google Gemini Capability

Google (2025) describes Gemini 2.5 Pro as the latest advancement in large language model technology. The model is designed to process multiple data formats, including text, code, and visual data. It can support complex reasoning tasks. The peak capability of the model is a large context window, capable of storing over one million tokens. This feature supports the model to analyze and summarize large datasets more efficiently than previous language models.

Another key feature of Gemini 2.5 Pro is its multilingual capability. Developers optimized the model to support more than 200 languages. In technical evaluations reported by Google, the model also demonstrated strong performance in tasks that involve language structure, factual consistency, and error identification. For example, Gemini scored 82.2% in the Aider Polyglot benchmark. It can imply that the model efficiently restructures content and detects linguistic errors in complex texts (Gemini Team, 2025).

With these technical strengths, Gemini has the potential to enhance English language learning for non-native speakers by providing effective AI-driven support. This section examines how AI improves efficiency in four key areas of instruction: learner motivation, personalized instruction, immediate feedback, and learning support (Al-Kadi & Ali, 2024). Specifically, Google Gemini acts as an analytical tool that can accurately identify students' writing patterns and knowledge gaps. Within a paragraph scoring framework, the model extracts key learning objectives from educational standards, ensuring that assessment feedback remains aligned with the curriculum while simplifying complex rhetorical concepts.

Thus, Gemini facilitates a diagnostic feedback cycle that goes beyond numerical scoring and supports personalized learning processes in academic writing (Rane et al., 2024).

When applied to the Thai education system, Google Gemini is highly recommended due to its in-depth analytical capabilities and alignment with the curriculum. These features are crucial for the long-term development of English as a Foreign Language (EFL) learners' skills. This analytical capability is particularly important for Thai students, who often struggle with English sentence construction and academic writing, as Gemini provides accurate feedback based on criteria tailored to their specific needs. Ultimately, integrating Gemini into the classroom will enable teachers to create a more personalized learning environment that effectively caters to the learning path of each individual Thai student (Kawinkoonlasate, 2025).

2.5 Prompt Design

Prompt engineering has recently become an important component in the effective use of Large Language Models (LLMs), particularly in educational applications. In general terms, prompt design refers to the process of structuring instructions given to a language model in order to guide the form and quality of its responses. Previous studies have emphasized the importance of designing prompts carefully so that model outputs remain consistent with instructional goals and assessment criteria (Anam, 2025; Correia et al., 2025).

Islam & Ahmed (2024) note that the effectiveness of LLM-based assessment systems is strongly influenced by the type of prompting strategy used. These strategies can be broadly grouped according to the level of contextual guidance provided to the model. Four commonly discussed prompting approaches are outlined below:

(1) Zero-Shot (0-shot) Prompting. In this approach, the model performs a task using only its existing training knowledge and a textual description of the task, without receiving explicit examples.
(2) Few-Shot (n-shot) Prompting. Here, the model is provided with several example input–output pairs that demonstrate the expected response format. These examples help the model interpret the task and produce outputs that better match the required structure.
(3) Chain-of-Thought (CoT) Prompting. This strategy encourages the model to generate intermediate reasoning steps before arriving at a final answer. By breaking down complex tasks into smaller reasoning processes, CoT prompting can improve transparency and accuracy in model responses.
(4) Majority Vote (maj1@k). In this method, multiple responses are generated for the same task, and the most frequently occurring output is selected as the final result. This approach can help improve reliability by reducing the influence of occasional model errors.

Research indicates that Gemini 2.5 can reach a level of scoring accuracy comparable to human raters when it is evaluated using basic, context-enhanced, and chain-of-thought (CoT) prompting strategies. With these strategies, the model can better manage complex linguistic differences in student writing. These strategies also promote assessment equity for English language learners (ELLs) by reducing disparities prevalent in traditional systems. Therefore, prompt design plays an important role in LLM-based assessment. Careful prompting helps maintain the validity and reliability of educational evaluations (Huang & Wilson, 2021). Ultimately, these advanced prompting protocols transform Gemini into an intelligent diagnostic instructor. Through this process, the system can provide transparent and

comprehensive feedback that helps students manage their learning more effectively (Sardi et al., 2025).

2.6 Validity and Reliability of AI Scoring Systems

The reliability of Automated Essay Scoring (AES) is primarily defined by its alignment with human evaluation (Weigle, 2002). Recent studies show that Large Language Models (LLMs) can achieve high-scoring consistency. Pearson correlation values between 0.70 and 0.85 have been reported, which are similar to human–human agreement levels (Mizumoto & Eguchi, 2023; Yancey et al., 2023). These models can match the performance of legacy systems such as PEG when context-enhanced few-shot chain-of-thought prompting is applied. They can also produce more balanced assessments for English language learners (Huang & Wilson, 2021).

Kawinkoonlasate (2025) described the use of Google Gemini in the Thai educational context. The study found that multimodal feedback and source integration can improve construct validity, especially for lower-proficiency learners. Despite this progress, there is still a notable lack of research focusing specifically on advanced reasoning models for Thai high school students. In particular, it is still unclear how effectively models such as Gemini can detect subtle communicative errors. This challenge is especially relevant within the regulatory framework of the Thai Basic Education Core Curriculum.

Previous literature indicates that the Thai Basic Education Core Curriculum requires analytic scoring to evaluate complex communication skills. This type of assessment is often time-consuming. At the same time, the Gemini 2.5 Pro model possesses advanced reasoning capabilities and reliability that may allow it to function as an automated assessment tool. However, empirical evidence supporting its application in Thai secondary schools remains limited. In particular, few studies have integrated this model with the Ministry of Education's specific requirements for assessing high school writing. Therefore, this study developed a system that integrates analytic scoring criteria, systematic prompt design, and the Gemini API to ensure curriculum-aligned assessment for Thai students. This gap highlights the need for further investigation into the effectiveness of advanced reasoning models in detecting subtle communicative features in Thai EFL writing.

3. Research Methodology

3.1 Research Design

This study employs a mixed-methods research design using an explanatory sequential framework. The quantitative phase assesses inter-rater reliability by using Quadratic Weighted Kappa (QWK) to measure the alignment between Gemini and human scores based on the Thai basic education core curriculum. Subsequently, the qualitative phase analyzes the AI-generated feedback to contextualize these statistical results, demonstrating the system's rigorous scoring logic and pedagogical value.

3.2 Participants

The population involved upper secondary EFL students enrolled in a Fundamental English course at a secondary school in Mahasarakham province, Thailand. In this study, cluster random sampling was employed to select participants from the available classrooms to ensure a representative range of writing proficiencies. A total of 160 participants were

selected for data collection because this sample size provides a sufficient dataset to statistically validate the consistency and reliability of the automated scoring system.

3.3 Instruments

3.3.1 Writing Tasks

The writing tasks were developed based on the Health unit of the Thai Basic Education Core Curriculum, focusing on the topic of sports and exercise. The tasks required students to produce expository paragraphs, which emphasize explaining reasons or motivating factors related to a particular topic. This paragraph type was selected because it aligns with the communicative objectives of Standards F1.2 and F1.3.

To elicit authentic student performance, three writing prompts were designed.

(1) Reasons why playing sports makes me happy
"Write a paragraph explaining the causes why playing sports makes you feel good. Start with a topic sentence that states how sports affect your mood, and then give specific reasons."

(2) My favorite sport
"Choose a sport you like the most. Write a paragraph explaining the reasons why this sport is your favorite. Use a clear topic sentence and support it with reasons such as the excitement of the game or the skills involved."

(3) What motivates me to exercise
"Write a paragraph explaining the factors that make you exercise or play sports regularly. Start with a topic sentence stating that you have a habit of exercising, and explain what drives this behavior."

3.3.2 Analytic Scoring Rubric

A criterion-referenced analytic rubric was developed to evaluate the writing samples. A panel of experts validated the rubric using the Index of Item–Objective Congruence (IOC). This step ensured that the rubric aligned with the objectives of the Thai Basic Education Core Curriculum.

The rubric included four assessment domains. Each domain was scored on a five-point scale, giving a total possible score of 20 points:

(1) Format & Content
(2) Organization & Coherence
(3) Sentence Construction & Grammar
(4) Vocabulary

3.3.3 Gemini API-Based Automated English Paragraph Scoring

The automated scoring instrument was a custom-developed application using JavaScript to interface with the Gemini 2.5 Pro API. The system was engineered with specific prompts designed to simulate an expert evaluator. These prompts instructed the model to generate a comprehensive assessment. The response contained three components: numerical scores based on the four-domain analytic rubric, a rationale explaining each score, and suggestions for improvement. Furthermore, the system was programmed to directly reference specific errors or strengths within the student's text.

3.4 Data Analysis Procedures

To address the research questions, the data analysis was divided into quantitative and qualitative components.

3.4.1 System Development and Qualitative Evaluation (RQ1)

The qualitative data for RQ1 consisted of system design processes and prompt development. The analysis focused on two aspects: (1) *System Architecture and User Interface.* This analysis examined how the JavaScript-based interface was integrated with the Gemini 2.5 Pro API. It focused on how the input system and output display supported the automated scoring process. (2) *Prompt Engineering Analysis.* This stage documented the iterative refinement of prompt structures to ensure that the generated outputs aligned with the analytic scoring rubric.

3.4.2 Quantitative Data Analysis (RQ2)

The quantitative data consisted of numerical scores assigned to students' writing by three human raters and the Gemini-based system. (1) *Inter-rater Reliability (ICC).* The Intraclass Correlation Coefficient (ICC) was used to examine the consistency among human raters. (2) *Descriptive Statistics.* Means, standard deviations, and frequency distributions were calculated to describe scoring patterns. (3) *Scoring Consistency (QWK).* Quadratic Weighted Kappa (QWK) was used to measure the level of agreement between human raters and the Gemini-based scoring system.

3.4.3 Qualitative Data Analysis (RQ2)

The qualitative data consisted of AI-generated feedback on students' writing. A subset of the feedback was analyzed using thematic analysis to examine whether the explanations provided by the system were consistent with the assigned scores and aligned with the curriculum-based writing criteria.

4. Results

4.1 Development of the Gemini API-Based System, Qualitative data (RQ1)

The development of this automated scoring tool focused on connecting the user interface smoothly with the Gemini 2.5 Pro engine. The following sections describe the system's design and its main parts. These include the overall architecture, the method used for prompt engineering, the design of the user interface, and how the results and feedback are shown.

4.1.1 System Architecture

The system operates as a client-side web application built with JavaScript, allowing it to run efficiently without needing complex server infrastructure. As shown in Figure 1, the process begins when the user submits a writing task through the interface. The application then creates a dynamic prompt by combining the student's input with the system's scoring rubrics. This data is sent to the Gemini API via a secure request. Finally, the system processes the returned JSON response to display the results in a clear, readable format for the user.

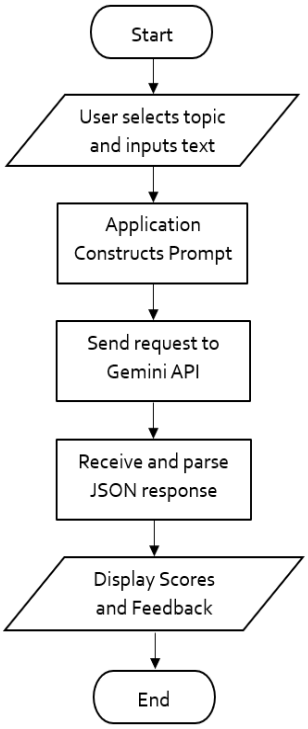

Figure 1: System Architecture and Data Flow

4.1.2 User Interface

The User Interface (UI) is built to be user-friendly for students. As shown in Figure 2, the input screen has a simple layout. It has a drop-down menu for picking a topic and a text field for input. This design makes it simple to submit writing tasks seamlessly.

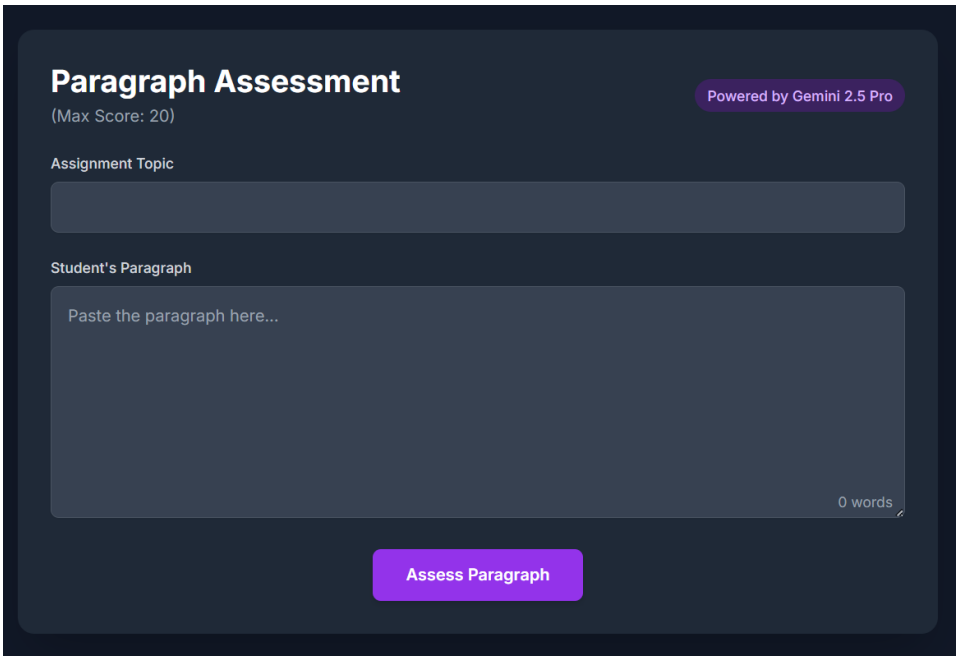

Figure 2: User Interface for Task Selection and Writing Input

4.1.3 Output Display and Feedback

As soon as the system gets the data from the Gemini API, it shows the results right away so that you may get immediate and valuable feedback. The output screen in Figure 3 has a complete Score Card. This shows how well the student did in four important categories, so they may quickly assess where they are. There is also a part that shows feedback in text form. In this section, the AI's advice and particular ideas are grouped by type. This helps right away by letting learners fix their own mistakes and get better on their own.

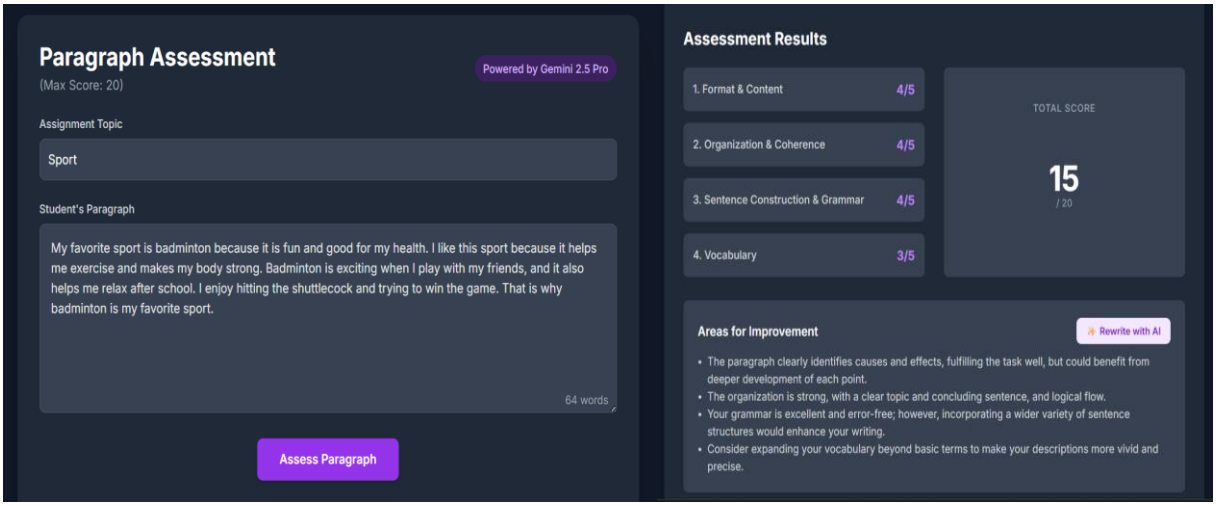

Figure 3: Display of Automated Scores and Diagnostic Feedback

4.1.4 Prompt Engineering

To ensure validity and reliability, the system uses a specific prompt engineering strategy designed to mimic expert human evaluation. This helps AI models become more consistent, addressing a common issue in automated grading. The prompt structure has three main parts. It starts by giving the model a specific task: to act as an experienced EFL teacher for Thai high school students. This ensures that the assessment is at the proper level of difficulty. The prompt also includes the full four-domain analytic rubric covering format & content, organization & coherence, sentence structure, and vocabulary to make sure that the grading is strict and standardized. Finally, the model is required to generate data in a JSON format. This ensures that all scores and comments are consistent, making it easy for the frontend system to process and display the results.

4.2 Performance Evaluation of the System, Quantitative data (RQ2)

To validate the developed system, its performance was rigorously evaluated against human judgments. This section presents the statistical findings regarding the reliability of human raters, the descriptive statistics of scoring patterns, the consistency between human and AI scores, and a qualitative analysis of the AI-generated feedback.

4.2.1 Reliability of Human Raters

To establish a reliable benchmark, the study first examined the consistency among the three human raters. The Intraclass Correlation Coefficient (ICC) was used to measure how strongly the experts agreed on the scores assigned to the students. This ensures that the averaged human scores used for comparison are accurate and trustworthy.

Table 3: Intraclass Correlation Coefficient (ICC) Indices of Inter-Rater Reliability among Three Human Raters

Domains	ICC Value	Interpretation
Format & Content	0.72	Moderate Reliability
Organization & Coherence	0.74	Moderate Reliability
Sentence Construction & Grammar	0.80	Good Reliability
Vocabulary	0.84	Good Reliability
Overall Average	0.78	Good Reliability

According to Table 3, the overall Intraclass Correlation Coefficient (ICC) was 0.78. Based on Koo & Li (2016), this shows a 'good' level of agreement among the three raters. When looking at specific areas, sentence construction & grammar (0.80) and vocabulary (0.84) showed high consistency. This suggests that experts can easily agree on objective linguistic errors. In contrast, the format & content (0.72) and organization & coherence (0.74) domains showed moderate reliability. This slight difference is natural, as judging the flow of ideas (discourse-level) is more subjective than checking simple mechanics. However, since the overall average stays above the 0.75 threshold, the human scores are statistically solid. They serve as a reliable benchmark for testing the automated system in this study.

4.2.2 Descriptive Statistics of Scoring Patterns

To check for potential scoring bias, the study compared the mean and standard deviation of the human ratings against the Gemini API scores across all 160 samples. The aim was to determine if the AI showed any significant severity or leniency.

Table 4: Comparison of Means and Standard Deviations between Gemini API and Human Raters

Domains (max score: 5)	Human (M)	Human (SD)	Gemini API (M)	Gemini API (SD)	Mean Diff.
Format & Content	3.52	0.85	3.45	0.82	-0.07
Organization & Coherence	3.15	0.90	3.08	0.91	-0.07
Sentence Construction & Grammar	3.10	1.10	2.95	1.05	-0.15
Vocabulary	3.40	0.80	3.35	0.78	-0.05
Total Score (max score: 20)	13.17	3.45	12.83	3.38	-0.34

According to Table 4, the Gemini API is slightly stricter than the human raters. The AI-generated scores were consistently lower across all domains, resulting in a total mean difference of -0.34. This trend was most noticeable in the sentence construction & grammar section, showing a difference of -0.15. This suggests that the AI follows linguistic rules strictly, whereas human raters might be more forgiving of minor errors. However, despite this stricter scoring, the standard deviations (SD) remained highly consistent between the two groups. This indicates that while the AI gives slightly lower scores, it effectively maintains the distinction between high and low-proficiency students, matching the human benchmark's distribution.

4.2.3 Consistency Analysis

To validate the system's reliability, the study examined the agreement between the Gemini API and human ratings using Quadratic Weighted Kappa (QWK). This metric was chosen because it penalizes major scoring discrepancies more heavily than slight variations, making it ideal for this type of data.

Table 5: Quadratic Weighted Kappa (QWK) Agreement Indices between Gemini API and Human Raters

Domains	QWK Value	Level of Agreement
Format & Content	0.81	Almost Perfect
Organization & Coherence	0.79	Substantial
Sentence Construction & Grammar	0.86	Almost Perfect
Vocabulary	0.82	Almost Perfect
Overall Average	0.82	Almost Perfect

Table 5 shows that the Gemini API aligns closely with the human benchmark. The overall Quadratic Weighted Kappa (QWK) score was 0.82, which indicates almost perfect agreement. The system performed effectively in the sentence construction & grammar domain, reaching a high of 0.86. Similarly, the vocabulary domain showed strong reliability with a value of 0.82. These figures confirm that the API's strict adherence to linguistic rules corresponds strongly with expert evaluation. On the macro level, the format & content domain achieved 0.81, while organization & coherence recorded a substantial 0.79. Since the overall QWK points to almost perfect agreement, the Gemini API stands out as a highly reliable tool for automated assessment in this context.

4.3 Gemini API Feedback, Qualitative data (RQ2)

To further validate the statistical results, this section presents a qualitative examination of the feedback generated by the Gemini API. Figure 4 illustrates a representative assessment case under the topic of sport, displaying the specific student submission:

'I like basketball the most because it is fun to play. This sport is exciting because players have to run and score points quickly. Basketball helps me exercise and makes me feel healthy. I enjoy playing it with my friends after school. That is why basketball is my favorite sport.'

This case serves to compare the student's input with the system's evaluation, thereby highlighting the practical application of the scoring criteria.

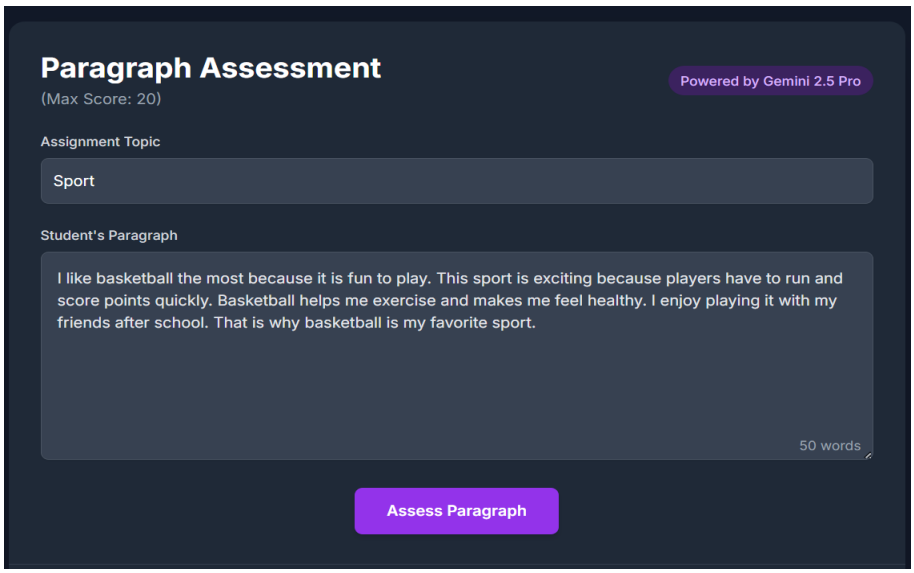

Figure 4: Illustrate a student paragraph

To provide a concrete illustration of the assessment process and to substantiate the statistical findings with qualitative evidence, a representative evaluation case is presented. Figure 5 shows the system's output interface for a student paragraph on the topic of sport, detailing the breakdown of scores across the four domains and the specific constructive feedback generated in the Areas for Improvement section.

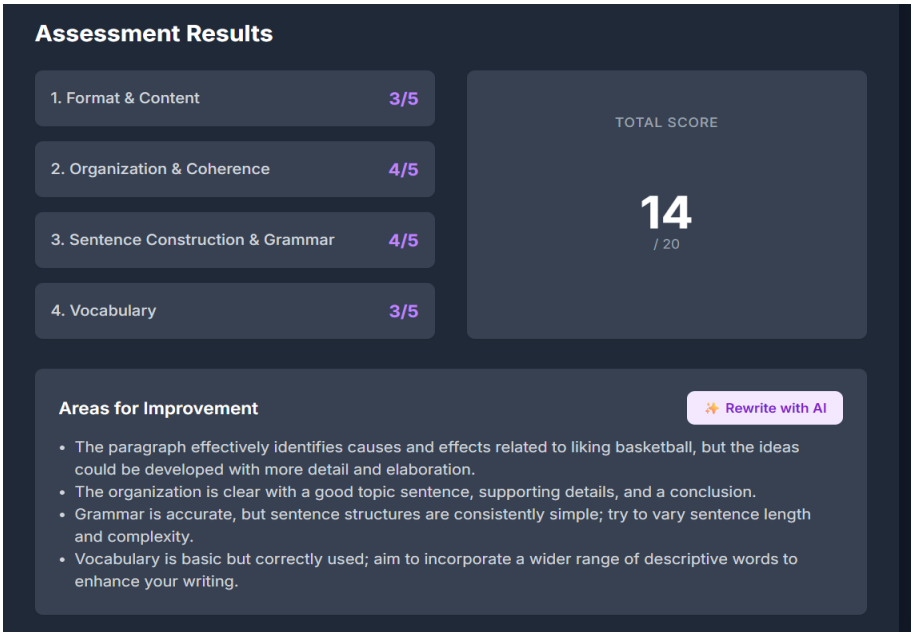

Figure 5: Illustrates a representative assessment case under the topic of sport

A qualitative analysis of the student's submission on the topic of sport, as detailed in Figure 5, substantiates the statistical findings regarding the Gemini API's rigorous scoring standards. The assessment highlights the system's capability to differentiate between functional correctness and advanced proficiency. For instance, despite the student submitting a grammatically error-free text with a clear topic, the system assigned a vocabulary score of 3/5 and a sentence construction & grammar score of 4/5, rather than perfect marks. The diagnostic feedback explicitly justified these ratings, noting that while grammar is accurate, the sentence structures are consistently simple, and similarly, that vocabulary is basic but correctly used. This confirms that the Gemini API evaluates beyond mere correctness, necessitating syntactic complexity and lexical variety for top-tier scores. Moreover, by advising the student to vary sentence length and incorporate a wider range of descriptive words, the system demonstrates its pedagogical value in offering actionable, formative guidance for learner improvement.

5. Discussion

5.1 Development of the Gemini API-Based System (RQ1)

Regarding the first objective, this study describes the development of an automated assessment system aligned with the policy of the Office of the Basic Education Commission (2025). The policy encourages the use of artificial intelligence to improve teacher efficiency and provide diagnostic feedback on student performance.

The effectiveness of the system can be explained by its structured three-part prompt framework. First, role definition positions the AI as an expert EFL teacher for Thai high school students. Second, rubric integration applies scoring criteria based on the Thai Basic Education Core Curriculum. Third, strict output constraints require the model to generate responses in JSON format.

These findings confirm and extend previous research by White et al. (2023) and Nguyen et al. (2025). Their studies emphasized that carefully structured prompts, especially those including clear role descriptions and scoring rubrics, are important for improving the reliability of generative AI systems. In line with these studies, the present research shows that structured prompting can guide large language models to perform more consistent evaluation tasks in EFL assessment contexts.

This study also contributes to recent technical developments in AI-assisted assessment. For example, Mughal et al. (2026) developed a system that used regular expressions (Regex) to extract scores from unstructured AI responses. In contrast, the system developed in this study applies strict JSON output constraints directly within the prompt. This approach helps reduce parsing errors caused by model verbosity. It also improves data integrity when transferring outputs from the Gemini API to the user interface.

In addition, the platform developed in this study differs from previous systems in its design. While Mughal et al. (2026) used a Flask-based graphical interface for flexible rubric structures, the present system uses a localized web-based design aligned with Thai educational assessment standards. This contextualized design addresses classroom challenges identified by Elhag et al. (2025), particularly the heavy grading workload faced by teachers. As a result, the system can simplify routine scoring tasks and allow teachers to focus more on instructional feedback and targeted support for student learning gaps.

5.2 Performance Evaluation of the System (RQ2)

Regarding the second research objective on system performance, the quantitative findings indicate a high level of reliability for the Gemini-based assessment system. The system achieved an overall Quadratic Weighted Kappa (QWK) of 0.82. According to the interpretation framework proposed by Landis & Koch (1977), this value represents an "almost perfect" level of agreement between the Gemini-generated scores and human raters. This result suggests that the automated scoring produced by the Gemini API is closely aligned with human evaluation. It also approaches the level of consistency typically expected between trained human raters.

The reliability observed in this study is higher than that reported in several recent studies on AI-assisted essay scoring. For example, Aydın et al. (2025) reported a QWK value of 0.72 when using a zero-shot prompting approach to evaluate L2 essays. The higher agreement found in the present study suggests that structured prompt engineering may improve scoring consistency. This is particularly evident when prompts include explicit role definitions and rubric-based constraints.

Similar results can be observed in studies focusing on Gemini-based scoring systems. Mughal et al. (2026) reported QWK values of approximately 0.43 to 0.45. Huang & Wilson (2021) also found only moderate agreement between Gemini-generated scores and human raters, with QWK values around 0.60. Compared with these findings, the higher reliability observed in the

present study suggests that rubric-aligned prompting and context-specific evaluation criteria may improve the scoring consistency of Gemini-based assessment systems.

These findings are also consistent with Kartika (2024), who reported that Gemini-based writing tools can effectively support key linguistic features such as grammatical accuracy, vocabulary use, and textual coherence. Taken together, the results suggest that Gemini, when supported by structured prompts and clearly defined scoring rubrics, can produce automated scores that closely align with human judgment in EFL writing assessment contexts.

5.3 Gemini API Feedback (RQ2)

Finally, the qualitative analysis of the AI-generated feedback highlights the pedagogical value of the Gemini-based system. This analysis is illustrated through the sports topic assessment. The feedback shows the system's ability to distinguish between functional correctness and linguistic complexity. It can award partial scores even when grammatical errors are minimal. This result suggests that the system can provide diagnostic feedback rather than only surface-level scoring.

This finding is consistent with Silva & Costa (2025), who reported that AI-generated feedback can help learners identify errors and understand weaknesses. It also supports the observations of Kasimova & Babakulova (2025), who noted that AI tools can provide timely and personalized feedback that supports writing development in EFL contexts. In addition, the specific suggestions generated by the system, such as varying sentence length and using more descriptive vocabulary, align with the findings of Trinh & Dan (2025), their study showed that students perceive Gemini as a useful assistant for improving writing mechanics and organization.

Furthermore, the present findings extend the work of Nguyen (2024) who suggested that Gemini-based learning environments can reduce learners' anxiety by providing a private and non-judgmental feedback space. Hou et al. (2024) also emphasized the value of personalized and on-demand feedback. Together, these results indicate that Gemini-generated feedback can support both diagnostic assessment and formative learning in EFL writing contexts.

Despite these promising findings, several limitations should be acknowledged. First, the study was conducted in a specific educational context involving Thai high school students. The rubric was also aligned with the Thai Basic Education Core Curriculum. Therefore, the findings may not be fully generalizable to other educational contexts or assessment frameworks. Second, the reliability analysis was based on a limited dataset of student essays and a small number of human raters. Although the results showed a high level of agreement between Gemini-generated scores and human evaluation, further research using larger and more diverse datasets is necessary. This would help confirm the robustness of the system across different writing tasks and learner populations. Finally, while the system demonstrated the potential to generate diagnostic feedback, the pedagogical usefulness of AI-generated feedback may still depend on prompt design and teacher mediation.

This study contributes to the field of writing assessment by demonstrating how a Gemini API-based system can support automated scoring and feedback in EFL writing contexts. Unlike many previous studies that focused mainly on general large language model performance, the present study provides empirical evidence that a structured prompt framework can

improve the reliability of AI-assisted scoring. The framework combines role definition, rubric integration, and output constraints. In addition, the findings highlight the pedagogical value of AI-generated feedback in supporting formative assessment. By providing immediate diagnostic feedback on linguistic features such as grammar, vocabulary, and organization, the system can help teachers identify students' writing difficulties. It can also encourage learners to revise and improve their writing.

6. Conclusion

This study aimed to develop and evaluate an automated English paragraph scoring system using the Gemini API, aligned with the writing indicators of the Thai upper secondary curriculum. The findings show that a structured prompt framework, combining role assignment, rubric integration, and strict output constraints, can support reliable automated scoring that closely aligns with human evaluation. The system demonstrated a high level of agreement with human assessors, particularly in grammatical accuracy, suggesting that well-designed prompt structures can improve the consistency of AI-assisted writing assessment in EFL contexts. In addition, the results highlight important pedagogical implications, as the system can provide diagnostic feedback on linguistic features such as grammar, vocabulary, and organization. This type of feedback can support teachers in identifying students' writing difficulties while encouraging learners to revise and improve their work.

However, this study was conducted within a specific educational context involving Thai high school students and was based on a limited dataset. These constraints may affect the generalizability of the findings. Therefore, future research should examine the applicability of similar AI-assisted assessment systems using larger datasets, a wider range of writing tasks, and more diverse educational contexts. Overall, this study contributes to the field of writing assessment by demonstrating how structured prompt design can support reliable and pedagogically meaningful AI-assisted writing evaluation.

References

Ajabshir, Z. F., & Ebadi, S. (2023). The effects of automatic writing evaluation and teacher-focused feedback on CALF measures and overall quality of L2 writing across different genres. *Asian-Pacific Journal of Second and Foreign Language Education*, *8*(1), 26. https://doi.org/10.1186/s40862-023-00201-9

Aliakbari, M., Barzan, P., & Allahveysi, S. P. (2025). AI usage in academic writing: Perspectives of stakeholders. *AI and Tech in Behavioral and Social Sciences*, *3*(4), 1–12. https://doi.org/10.61838/kman.aitech.4343

Al-Kadi, A., & Ali, J. K. M. (2024). A holistic approach to ChatGPT, Gemini, and Copilot in English learning and teaching. *Language Teaching Research Quarterly*, *43*, 155–166. https://doi.org/10.32038/ltrq.2024.43.09

Anam, R. K. (2025). Prompt engineering and the effectiveness of large language models in enhancing human productivity. *ArXiv*. https://doi.org/10.31219/osf.io/ad9y5_v1

Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, *15*(1), 14551. https://doi.org/10.1038/s41598-025-87862-3

Aydın, B., Kışla, T., Elmas, N. T., & Bulut, O. (2025). Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System*, *133*, 103784. https://doi.org/10.1016/j.system.2025.103784

Correia, A. P., Hickey, S., & Xu, F. (2025). Realizing the possibilities of the large language models: Strategies for prompt engineering in educational inquiries. *Theory Into Practice*, *64*(4), 434–447. https://doi.org/10.1080/00405841.2025.2528545

Elhag, A., Al Abri, M., & Yousef, A. M. F. (2025). The effect of generative AI tools (ChatGPT, Gemini, etc.) on students' achievement and their motivation towards learning. *Journal of Technology and Science Education*, *15*(3), 746. https://doi.org/10.3926/jotse.3410

Firoozi, T., Mohammadi, H., & Gierl, M. J. (2023). Using active learning methods to strategically select essays for automated scoring. *Educational Measurement: Issues and Practice*, *42*(1), 34–43. https://doi.org/10.1111/emip.12537

Gemini Team, G., Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., … Helmholz, W. (2025). *Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities*. http://arxiv.org/abs/2507.06261

Google. (2025). *Gemini 2.5 Pro Model Card*. Https://Storage.Googleapis.Com/Deepmind-Media/Model-Cards/Gemini-2-5-Pro-Model-Card.Pdf.

Harmer, J. (2007). *How to teach English*. Pearson Longman.

Harmer, J. (2015). *The practice of English language teaching*. Pearson Longman.

Hou, X., He, S., & Cuigong, R. (2024). Learner use of AI-generated feedback for written corrective feedback in L2 writing: usefulness, user proficiency, and attitude. *Proceedings of the 2024 8th International Conference on Education and Multimedia Technology*, 70–76. https://doi.org/10.1145/3678726.3678767

Huang, Y., & Wilson, J. (2021). Using automated feedback to develop writing proficiency. *Computers and Composition*, *62*, 102675. https://doi.org/https://doi.org/10.1016/j.compcom.2021.102675

Hyland, K., & Richards, J. C. (2014). *Second language writing*. Cambridge University Press.

Imran, M., & Almusharraf, N. (2024). Google Gemini as a next-generation AI educational tool: a review of emerging educational technology. *Smart Learning Environments*, *11*(1), 22. https://doi.org/10.1186/s40561-024-00310-z

Islam, R., & Ahmed, I. (2024). Gemini-the most powerful LLM: Myth or truth. *2024 5th Information Communication Technologies Conference (ICTC)*, 303–308. https://doi.org/10.1109/ICTC61510.2024.10602253

Jarunthawatchai, W., & Baker, W. (2024). English language education and educational policy in Thailand. In *The Oxford Handbook of Southeast Asian Englishes* (pp. 557–574). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780192855282.013.30

Kartika, S. (2024). Enhancing Writing Proficiency through AI-Powered Feedback: A Quasi-Experimental Study Using Google Gemini. *LinguaEducare: Journal of English and Linguistic Studies*, *1*(2), 83–96. https://doi.org/10.63324/h6q1ak58

Kasimova, M., & Babakulova, D. (2025). A comparative evaluation of ChatGPT, Gemini, and Perplexity feedback for B1-B2 EFL learners. *Foreign Languages in Uzbekistan*. https://doi.org/10.36078/1767688041

Kawinkoonlasate, P. (2025). A comparative study of Google Gemini and ChatGPT in enhancing English language learning for EFL learners: A case study of the English

research writing course. *Pedagogical Research*, *10*(4), em0251. https://doi.org/10.29333/pr/17670

Kim, Y., Mozer, R., Al-Adeimi, S., & Miratrix, L. (2025). *ChatGPT vs. Machine Learning: Assessing the efficacy and accuracy of large language models for automated essay scoring.* (EdWorkingPaper:25-1335). Annenberg Institute at Brown University. https://doi.org/https://doi.org/10.26300/7vj9-5y53

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Langan, J., & Albright, Z. L. (2020). *Exploring writing: paragraphs and essays*. McGraw-Hill Education.

Ministry of Education. (2008). *The Basic Education Core Curriculum B.E. 2551*. The Agricultural Co-operative Federation of Thailand, Ltd. Press.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mughal, N., Imran, A. S., Daudpota, S. M., Kastrati, Z., & Noor, W. (2026). Exploring the potential of large language models for automated essay scoring in education. *Discover Artificial Intelligence*, *6*(1), 166. https://doi.org/10.1007/s44163-026-01002-y

Nguyen, D. L., Le, P. T. T., & Le, T. T. (2025). Using Gemini for formative assessment in English academic writing - critical insights into the ai tool's efficacy. *AsiaCALL Online Journal*, *16*(1), 328–343. https://doi.org/10.54855/acoj.2516117

Nguyen, P. D. A. (2024). Gemini Google: A potential tool for English learning. *Thu Dau Mot University Journal of Science*, *6*(3). https://doi.org/10.37550/tdmu.EJS/2024.03.586

Office of the Basic Education Commission. (2025). *AI usage guide for teachers, students, schools, and parents in Thailand*. Bureau of Academic Affairs and Educational Standards.

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*, 100020. https://doi.org/10.1016/j.caeai.2021.100020

Rane, N., Choudhary, S., & Rane, J. (2024). Gemini versus ChatGPT: Applications, performance, architecture, capabilities, and implementation. *Journal of Applied Artificial Intelligence*, *5*(1), 69–93. https://doi.org/10.48185/jaai.v5i1.1052

Sardi, J., Darmansyah, C., O., Yanto, D. T. P., & Eliza, F. (2025). How does generative AI influence students' self-regulated learning and critical thinking skills? A systematic review. *International Journal of Engineering Pedagogy (IJEP)*, *15*(1), 94–108. https://doi.org/10.3991/ijep.v15i1.53379

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge/Taylor & Francis Group.

Silva, P., & Costa, E. (2025). Assessing large language models for automated feedback generation in learning programming problem-solving. *ArXiv E-Prints*, arXiv:2503.14630. https://doi.org/10.48550/arXiv.2503.14630

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence.

Computers and Education: Artificial Intelligence, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Teng, M. F., Qin, C., & Wang, C. (2022). Validation of metacognitive academic writing strategies and the predictive effects on academic writing performance in a foreign language context. *Metacognition and Learning*, *17*(1), 167–190. https://doi.org/10.1007/s11409-021-09278-4

Tesfay, H. (2017). Investigating the practices of assessment methods in Amharic language writing skill context: The case of selected higher education in Ethiopia. *Educational Research and Reviews*, *12*(8), 488–493. https://doi.org/10.5897/ERR2017.3169

Tongsilp, A., Tangdhanakanond, K., & Chaimangkol, N. (2024). Development of automated scoring system for Thai writing ability test of primary education level. *Kasetsart Journal of Social Sciences*, *45*(3). https://doi.org/10.34044/j.kjss.2024.45.3.05

Trinh, N. T. N., & Dan, T. C. (2025). EFL students' perceptions and practices of using Gemini for developing English argumentative essay writing skills. *European Journal of Alternative Education Studies*, *10*(3). https://doi.org/10.46827/ejae.v10i3.6428

Vimala, A. (2025). English language learning in Thailand: Policy, practice, and pedagogy teaching approaches & methodologies. *Journal of Asian Language Teaching and Learning*, *6*(2). https://so10.tci-thaijo.org/index.php/jote/article/view/2986

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1249991

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2302.11382

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short l2 essays on the CEFR scale with GPT-4. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA, 2023)* (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49