**Research Journal of Psychology (RJP)**

**Online ISSN:** 3006-7219
**Print ISSN:** 3006-7200
Volume 3, Number 2, 2025, Pages 571 – 589
**Journal Home Page**
**https://ctrjournal.com/index.php/19/index**

# Evaluating Elementary-Level English Essays: Human-AI Synergy and the Role of Cognitive Load Theory

**Dr. Mamona Yasmin Khan[1], Shazia Riaz Cheema[2] & Sobia Tasneem[3]**

[1]*Professor of English, The Women University Multan, Email: mamona.6231@wum.edu.pk*
[2]*PhD Scholar, Department of English, The Women University Multan, Email: shazianazar78@gmail.com*
[3]*Lecturer of Education, National University Multan (NUML), Email: sobia.tasneem@numl.edu.pk*

**ARTICLE INFO**

***Article History:***

| | | |
|---|---|---|
| *Received:* | *April* | *21, 2025* |
| *Revised:* | *May* | *25, 2025* |
| *Accepted:* | *June* | *02, 2025* |
| *Available Online:* | *June* | *14, 2025* |

***Keywords:***

*Artificial Intelligence (AI); Human Evaluation; Cognitive Load Theory; Hybrid Grading System; Elementary-Level Writing Assessment*

***Corresponding Author:***
*Dr. Mamona Yasmin Khan*
***Email:***
*mamona.6231@wum.edu.pk*

## ABSTRACT

*This study examines the synergy between human evaluators and an AI-based system in assessing elementary-level English essays, focusing on key linguistic features such as grammar, syntax, spelling, content, and clarity. A dataset of 30 student-written essays is used to evaluate the effectiveness, reliability, and subjectivity of both evaluation methods. The boxplot comparing Human and AI evaluation scores offers insights into the evaluator's scoring behaviours in applying the grading rubric. Cronbach's Alpha values indicate high internal consistency in both evaluation methods, with human evaluators demonstrating slightly greater reliability. The study also integrates Cognitive Load Theory (Sweller, 1988) to explain the cognitive demands of human evaluators versus the rule-based processing of AI. These findings suggest that while AI provides efficiency in mechanical assessments, human evaluators bring a nuanced understanding, emphasising the complementary roles of both in educational assessment. The study advocates for a hybrid approach that combines the strengths of both human and AI evaluations to enhance assessment fairness and accuracy.*

## Introduction

Artificial Intelligence (AI) has been witnessing rapid advancements in recent times, with its applications being experienced in all sectors, including education. Certainly, one of the most promising uses in education to date is to have AI assist or even supersede the traditional human assessment of the student's written assignments, such as essays (Wu et al., 2022). More and more, schools and universities all around the world have adopted AI-powered tools like automated essay scoring (AES) systems, which have the potential to not only accelerate but also become more efficient and objective in carrying out the grade (Attali & Burstein, 2006). Nevertheless, the involvement of AI in educational assessment has received a heated reaction, especially when it concerns the evaluation of elementary-level writing. In this case, the subjectivity of the human

judge and the capacity of the person to discern subtle details from the work of the student, is an important instrument in the function of an implicit complete evaluation (Familoni, 2024).

**Background of Study**

There is a significant gap in our understanding of how human evaluators and AI systems assess key linguistic features such as grammar, vocabulary, coherence, and overall essay structure. In addition to using established linguistic criteria, human evaluators of essays (especially, those with experience in language instruction) base their judgements on how well the writing meets their professional experience and the contextual nuances of the text (Cumming, 2001). It helps them to be able to pick up on subtle differences in writing that could make the difference in determining whether that student wrote a well-rounded judgment or they were less coherent and didn't display creativity and individual expression. On the other hand, most AI systems are built using data-driven algorithms, that is, algorithms largely focus on these types of superficial aspects such as grammar, spelling and structure of a sentence. These systems perform admirably, always doing so to ensure that these mechanical elements are consistently evaluated (Dergaa et al, 2023; Firdaus et al., 2022), but they struggle and do not cover deeper or more nuanced areas of writing like writer's expression, style and thematic richness (Shermis and Burstein, 2013). This highlights the weakness of state-of-the-art AI models that fall short of the level of depth of human understanding, which is needed for thorough evaluation of a piece of writing.

**Significance of Study**

The study provides information about how AI systems work for educational assessment of elementary-level English essays, while researchers continue expanding this field. It reveals that AI demonstrates potential for the automation of grading standard item sets such as grammar and syntax, but it faces constraints when assessing subjective elements that humans can better perceive because of their ability to establish meaning and contextual understanding (Huang, 2024).

This research enables substantial understanding of AI grading strengths and weaknesses by analysing AI and human assessments of poem summaries. Bennett (2015) confirms that proper student evaluation needs AI to work alongside human judgment for achieving fair and accurate results, according to educators and policymakers.

**Research Objectives**

1. To compare the grading patterns of AI and human evaluators across key linguistic features.
2. To analyse the cognitive load experienced by both AI and human evaluators in assessing structured versus creative aspects of student writing.
3. To investigate the feasibility and potential benefits of implementing a hybrid grading system, where AI assists in evaluating structural elements while human evaluators focus on creativity and content depth.

**Research Questions**

4. How do AI and human evaluators differ in their approach to grading key linguistic features, including grammar, syntax, spelling, content, clarity, and creativity?
1. What cognitive load factors contribute to the differences observed between AI and human evaluations, especially in creative or complex writing tasks?

2. How can AI and human evaluators work together, focusing on structural elements and human evaluators assessing creativity and content depth?

## Literature Review

The incorporation of Artificial Intelligence (AI) into educational assessment, particularly for evaluating elementary-level English writing, is reshaping traditional evaluation methods and redefining teacher roles. Current literature emphasises the potential of AI to augment human evaluators by enhancing scoring consistency, delivering personalised feedback, and supporting individual learning needs (Aghaziarati, 2023; Rios-Campos et al., 2023). Tools such as automated essay scoring systems and intelligent writing assistants have demonstrated effectiveness in early education environments, although issues related to data security, algorithmic fairness, and contextual sensitivity remain unresolved (Mohammed, 2023).

Despite notable advancements, there is a critical shortage of research exploring the use of AI specifically within elementary writing assessment, a stage where cognitive development significantly influences performance (Huang, 2024). Addressing this gap is essential, particularly when considering the cognitive demands placed on young learners during both writing and feedback reception.

The role of educators is pivotal in mediating AI's impact. Empirical studies have demonstrated that professional training related to AI literacy enhances teachers' capacities to integrate AI tools to the extent in terms of assessment practices (Park & Kwon, 2023; Kim & Kwon, 2024). Alshehri (2023) emphasises that AI-based teaching strategies should be adapted for a better fit to the developmental standards and pedagogical standards. However, ethical issues such as transparency in the decision-making that AI uses, insufficient accountability and possible biases continue to remain unresolved, hindering AI-assisted evaluation credibility and its acceptance (Holmes et al., 2021; Yu & Yu, 2023).

Furthermore, AI's widespread use in educational assessment has been impeded by the fact that technological infrastructures of access to AI, as well as degrees of digital preparedness, differ across educators and students (Familoni, 2024). This also applies to evaluating elementary-level writing, though for writing, we must pay careful attention to the cognitive burden imposed by the feedback mechanisms used by AI. Applied to Cognitive load theory, AI-generated feedback would provide valuable insight into how feedback can be built to create more examples of cognitive load that support rather than hinder development in students' writing.

Overall, to maximise the benefits of AI-human collaboration in elementary essay evaluation, the understanding and crafting need to continue. By addressing it, such efforts are critical to making sure that integration of AI in the areas of education and cognitive skills continue to achieve both educational and cognitive needs of young learners and promoting fairness, transparency and inclusivity (Mane, 2025).

### Human vs. AI Evaluation in Writing Assessment

The more traditional type of conventional writing assessment has increasingly been transformed by artificial intelligence (AI) from the traditional manually administered evaluations to more dynamic and data-driven evaluations. This paper aims to analyse the comparative advantages and

disadvantages of the human and AI-assisted writing evaluation, and applies them to the educational setting, especially in early language instruction.

The writing assessment tools that use AI (Automated Writing Evaluation systems and Automated Essay Scoring ones) rely on natural language processing (NLP) and machine learning algorithms to evaluate such writing elements as grammar, structure, coherence, etc. Alharbi (2023) traces the evolution of these tools from simple grammar checkers to sophisticated platforms capable of delivering comprehensive, real-time feedback. These technologies are especially useful in formative assessment contexts of writing, providing immediate and scalable answers to help students continuously improve their writing. This view is also supported by Wale (2024), who states that AI-based diagnostics can be useful to compensate for the lack of traditional instruction as it aims to mitigate mechanical, organisational, and linguistic weaknesses in student writing.

But the issue is that no one can confirm if that feedback is valid and reliable if generated by AI. AI tools excel at processing and, to some extent, evaluating a massive amount of text, but they struggle to identify when it has been written poorly, sounds or reads insipidly, or lacks adequate depth of thought and originality. Dergaa et al. (2023) claim that AI-generated feedback could encourage formulaic writing and doesn't create an element of engagement like strong academic prose. In addition, Wang (2024) suggests that AI literacy should be taught into the curriculum because students must be taught how to critically interpret AI feedback as opposed to passively trusting AI feedback. The danger of unguarded pedagogical practices is that they will not be good for students' critical thinking and creativity.

However, it need not be seen as contradictory that in this world, AI and human evaluation oppose each other. The best pedagogical framework may rather involve a combination of both approaches. Building upon this, Joo (2024) develops a hybrid model that coopts AI-generated feedback for preliminary analysis, but that relies on human evaluators for interpretive judgment, in more complex or subject areas such as argument strength, tone, or originality. Not only does this dual-layered approach provide for a more complete assessment, but it also allows for metacognitive development through comparison, reconciliation and answering of different types of feedback (Xie et al. 2025).

The use of AI in writing assessment also has an ethical aspect to it. The existing use cases (of AI-generated content) include misuse, including plagiarism, overreliance on AI-generated content, and academic integrity breaches are well documented. According to Perkins (2023), generative AI tools such as Chatgpt can work in facilitating learning when used ethically, but also can present serious risks of dodging writing. Therefore, educational institutions should develop a set of clear policies governing AI use in the institutions as an assistive tool rather than a substitute for the original works of students.

**Effective Feedback in Writing Assessment**

The more traditional type of conventional writing assessment has increasingly been transformed by artificial intelligence (AI) from the traditional manually administered evaluations to more dynamic and data-driven evaluations. This paper aims to analyse the comparative advantages and disadvantages of the human and AI-assisted writing evaluation, and applies them to the educational setting, especially in early language instruction.

The writing assessment tools that use AI (Automated Writing Evaluation systems and Automated Essay Scoring ones) rely on natural language processing (NLP) and machine learning algorithms

to evaluate such writing elements as grammar, structure, coherence, etc. Alharbi (2023) traces the evolution of these tools from simple grammar checkers to sophisticated platforms capable of delivering comprehensive, real-time feedback. These technologies are especially useful in formative assessment contexts of writing, providing immediate and scalable answers to help students continuously improve their writing. This view is also supported by Wale (2024), who states that AI-based diagnostics can be useful to compensate for the lack of traditional instruction as it aims to mitigate mechanical, organisational, and linguistic weaknesses in student writing.

But the issue is that no one can confirm if that feedback is valid and reliable if generated by AI. AI tools excel at processing and, to some extent, evaluating a massive amount of text, but they struggle to identify when it has been written poorly, sounds or reads insipidly, or lacks adequate depth of thought and originality. Dergaa et al. (2023) claim that AI-generated feedback could encourage formulaic writing and doesn't create an element of engagement like strong academic prose. In addition, Wang (2024) suggests that AI literacy should be taught into the curriculum because students must be taught how to critically interpret AI feedback as opposed to passively trusting AI feedback. The danger of unguarded pedagogical practices is that they will not be good for students' critical thinking and creativity.

However, AI and human evaluation oppose each other. The best pedagogical framework may rather involve a combination of both approaches. Building upon this, Joo (2024) develops a hybrid model that coopts AI-generated feedback for preliminary analysis, but that relies on human evaluators for interpretive judgment, in more complex or subject areas such as argument strength, tone, or originality. Not only does this dual-layered approach provide for a more complete assessment, but it also allows for metacognitive development through comparison, reconciliation and answering of different types of feedback (Xie et al. 2025).

The use of AI in writing assessment also has an ethical aspect to it. The existing use cases (of AI-generated content) include misuse, including plagiarism, overreliance on AI-generated content, and academic integrity breaches are well documented. According to Perkins (2023), generative AI tools such as Chatgpt can work in facilitating learning when used ethically, but also can present serious risks of dodging writing. Therefore, educational institutions should develop a set of clear policies governing AI use in the institutions as an assistive tool rather than a substitute for the original works of students.

**Theoretical Framework**

Cognitive Load Theory (CLT), initially introduced by Sweller (1988), offers a robust framework for understanding the cognitive challenges learners face when processing new information, particularly within the context of complex tasks that tax their working memory. CLT distinguishes between three types of cognitive load

- *Intrinsic load* (the inherent complexity of the material being learned).
- *Extraneous load* (the cognitive burden imposed by the manner of content presentation).
- *Germane load* (the mental effort invested in schema construction and meaningful learning).

Viewed through the lens of CLT, the cognitive demands on learners and evaluators are analysed for what they are, particularly when it comes to the complexity of writing tasks and the presentation of feedback.

CLT has immense value within elementary English essay evaluation, where it is used to distinguish between cognitive load observed because of the complexity of writing itself (intrinsic load) and that observed because of unclear or poorly formed feedback or prompts (extraneous load). Specifically, AI-driven assessment systems that use natural language processing (NLP) and machine learning algorithms have been shown to reduce extraneous cognitive load. These systems evaluate a set of essays quickly, structure the evaluation, and evaluate grammar, syntax, organisation, and adherence to linguistic conventions. Therefore, the use of AI in young learners helps to unload unnecessary cognitive burden by quickly and objectively providing feedback about where in a piece of writing the learner needs to improve structure or language mechanics through practice.

But AI systems are unable to appraise upper–order cognitive abilities that require subtlety of judgment, including creativity, abstract reasoning and the explication of complex ideas. Thus, AI pays greater attention to the more objective, rule-based parts of writing, which can be easily quantified: syntax, grammar, and organisation. This approach reduces extraneous load, but does not, in any similar way, address the mental effort that is inherent in producing imaginative, original, or non-standard written work. Conversely, humans as evaluators are better able to evaluate these types of complex dimensions. What they have in abundance is the refined judgment to gauge the level of depth in students' abstract reasoning, creativity and originality. In addition, human evaluators are in a better position to provide more holistic and contextually rich feedback about creative writing cues that reflect the cognitive complexities of creative writing.

A discussion is provided of Sweller's (1988) theory, which suggests the complementary strengths of human and AI evaluators in the assessment process. AI is good at managing the extraneous cognitive load as it caters for the structured and measurable aspects of writing, while the human evaluators are in a better place for addressing the intrinsic cognitive load on conceptualising and creative writing. This is why the synergy here, in elementary education, matters a lot, particularly in terms of building the dexterity alongside the originality. AI integration into the assessment process lets human evaluators focus on creativity and higher-order cognitive engagement by introducing more efficient management of extraneous load.

CLT can be integrated further into the design of the assessment to optimise the efficacy of human-AI collaboration in writing assessments. The AI feedback system should be engineered to minimise extraneous load by preferring clarity, breaking the feedback down into smaller chunks that could be digested, and avoiding cognitive overload. In turn, human evaluators can then concentrate on helping students scaffold other, higher-order writing skills, such as working through intrinsic cognitive requirements through model texts, guided questioning, and formative discussion strategies. When the assessment strategies are designed in such a way that they complement CLT principles, educators and developers of the same can create a balanced and conducive evaluation system that is not only efficient at delivering feedback but also supports the deeper learning outcomes by acknowledging the capacity limitations of young writers and creative expression. This integrated approach fosters a more comprehensive and effective writing assessment environment.

## Methodology

This study serves as an exploration of the human vs. artificial intelligence in sample measuring 30 mini-English essays of elementary school children using a mixed methods approach. The beginning is on various linguistic elements, such as content inclusion, grammar, sentence structure, coherence, and creativity, with the criterion outlined by the rubric developed. In this evaluation

process, the performance of the AI system is compared to that of the human evaluators in evaluating these key linguistic features from the perspective of the Cognitive Load theory (Sweller, 1988). By providing this theoretical perspective, the different cognitive loads of the human evaluators who, in essence, interpret and contextualise, and the AI system using rules-based processing are elucidated.

**Data Collection**

The dataset consists of 30 mini essays written by elementary school students, each summarising roughly 100-150 words of the poem '*A Time to Talk'* by Robert Frost. Human scoring and the Magic School AI assessment software were used to evaluate these essays twice. The study aims to compare these evaluations to determine the effect of cognitive load on the accuracy, efficiency, and subjectivity of both human and AI assessments.

**Assessment Rubric**

A comprehensive assessment rubric was used for evaluating elementary grade-level English essays. This rubric has been adopted from well-recognised standards both in the educational field as well as the AI field (IELTS, 2023; NWP, 2019). These rubrics are typically used to grade such facets of writing as grammar, syntax, clarity, content, etc., to ensure that there is fair and uniform grading of the same material. Scores were given on a scale of 0 to 3 each for the features, except for spelling, which was scored on a scale of 0 to 1; the rubric they should rely on reflects the following features:

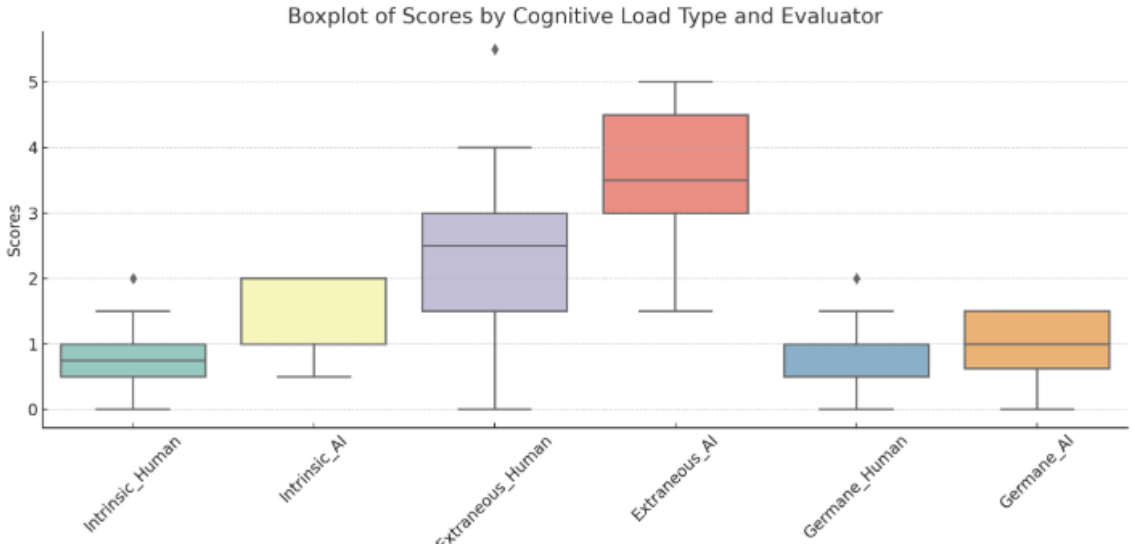
| Criteria | Description | Score Range |
|---|---|---|
| **Grammar** | Assesses the use of grammatical structures, including sentence agreement, verb forms, and tense consistency. | 0-3 |
| **Syntax/Sentence Structure** | Evaluates the logical arrangement, coherence, and complexity of sentences. | 0-2 |
| **Spelling** | Focuses on the accuracy of spelling and adherence to standard English conventions. | 0-1 |
| **Content** | Measures the relevance, completeness, and inclusion of key ideas aligned with the given topic. | 0-2 |
| **Clarity & Style** | Considers readability, coherence, and the use of descriptive or creative language. | 0-2 |

## Data Analysis

### 1. Quantitative Analysis

The Quantitative Analysis section is structured to closely relate comparative boxplot analysis with the following reliability testing in an order so that it is clear and logical. The score distribution of human and AI was observed to differ specifically in writing criteria through a comparative boxplot

analysis. Mechanical aspects of AI evaluators, i.e. grammar, syntax, and spelling, were shown to score on average higher and with more consistency than human evaluators, indicating their strength in low-cognitive load tasks, as in Cognitive Load Theory (CLT). These interquartile ranges are narrow, indicating high reliability and rule-based consistency. Human evaluators, however, exhibited much greater variability in clarity and content, which in the case of subjective judgment and contextual interpretation, are areas with large subjectivity and contextual interpretation. These findings confirm what intuitively should be true, that germane cognitive load, or load linked to the construction of schema and determination of mental model, is one more effectively handled by human reviewers. But AI also always gave higher total scores than human graders, and though they were better than human graders at score total, they did not produce the nuanced differentiation that you see in human graders, for example, in high load dimensions like creativity and message clarity.




The boxplot comparing Human and AI evaluation scores across the three core dimensions of Cognitive Load Theory (CLT), Intrinsic, Extraneous, and Germane Load, provides a visual insight into their respective scoring behaviours and assessment capabilities.

**1. Extraneous Cognitive Load (Grammar, Syntax, Spelling)**

Mechanical aspects of writing were also consistently better and more tightly clustered, and therefore more reliable and consistent when evaluated with AI scores. That is because AI is very good at doing rule-based assessments, such as grammar and spelling, which are not very cognitively demanding and do not require much thinking; they can be processed algorithmically. The human scores in this category were lower and more dispersed, consistent with inconsistency, either due to individual judgment differences or lack of errors overlooked. Less variance is shown in the narrower box for AI. AI is fatigue-free and unbiased subjectively for all samples.

**2. Intrinsic Cognitive Load (Content/Idea Complexity)**

Whereas AI scores were generally higher on average and exhibited more variation on extraneous tasks, content tended to display greater variability than AI. This implies that while AI attributes credit for structured or key term writing, it may overstate quality because it does not consider contextual nuances. In contrast, human scores in this category were lower and more measured,

indicating that scores tend to be higher when raters assess the complexity and coherence of ideas. The variability in different assessors' abilities to articulate how content quality and cognitive demand were perceived is broader than suggested by the overall human score range.

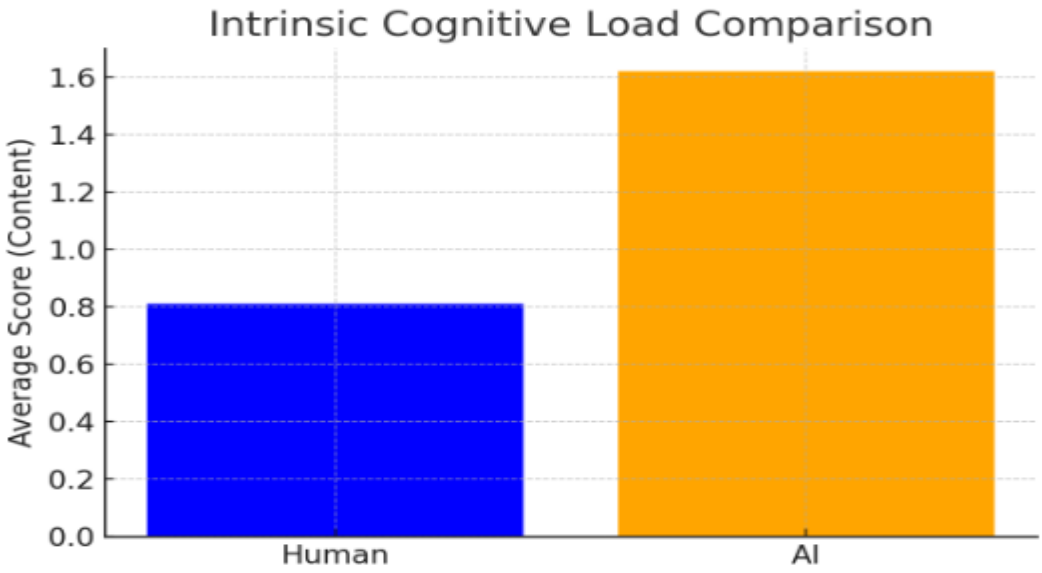
**3. Germane Cognitive Load (Clarity and Organisation)**

Median scores and the width of the distribution thus showed that human evaluators were again more sensitive to the clarity of the message and its coherence, as well as to the creative structure of the message. This is a type of load where humans will have to infer the understanding from the context by interpreting the intent, tone and flow, which is closed-loop understanding, and AI is yet to master the ability of measuring these effectively. The scores were lower and much more tightly concentrated, validating the view that algorithmic tools do not do as well at coping with tasks requiring the interpretation as well as the construction of schemata.

**Cognitive Load Theory Perspective:**

For all these reasons, the visual reinforced by the boxplot is that AI's forte is in the ability to reduce and evaluate the extraneous load, but it is less adept at intrinsic and germane cognitive load. While evaluators do display variability in their scores, human evaluators are more able to judge the complex and creative facets of student writing. It buttresses our point that human-AI synergy, in concert with CLT principles, can result in writing assessment systems with more balanced and comprehensive writing assessments that are particularly apt for elementary youngsters whose writing usually integrates technical, creative, and developmental factors.

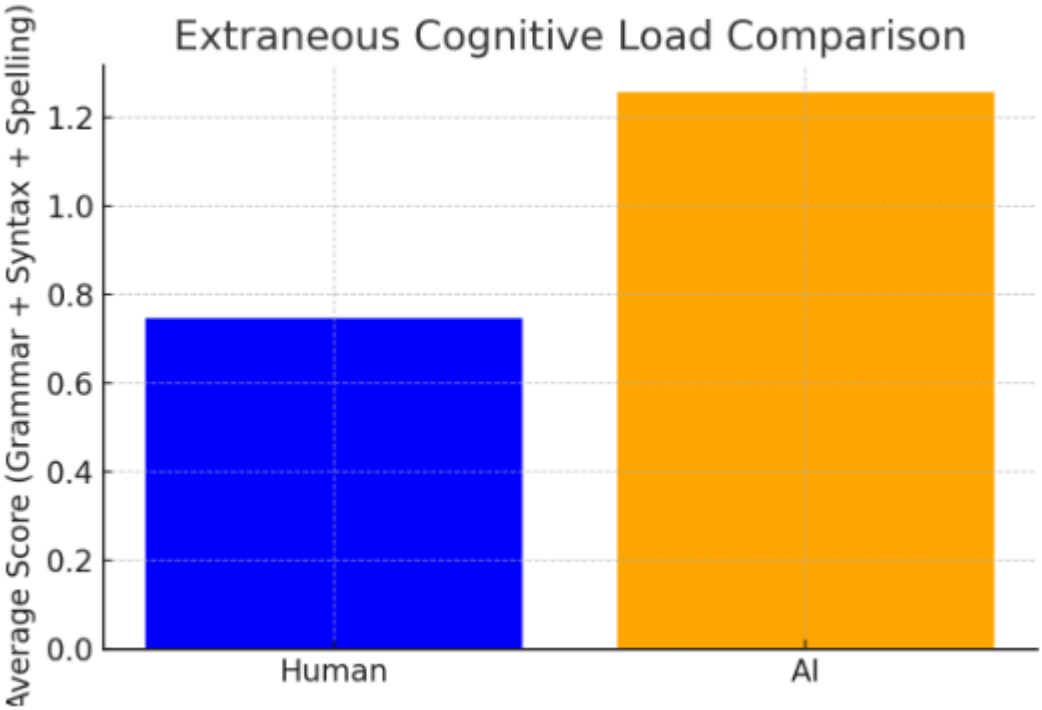
The boxplot visually confirms what AI can and cannot do to reduce and evaluate extraneous load, but it is inadequate at coping with intrinsic and germane cognitive load. Despite the variability of human scores, human evaluators can assess more complex, more creative, and more interpretive aspects of student writing. This aligns with the fundamental idea that if humans and AI can work together and align with CLT principles, it would contribute to writing assessment systems that are more balanced and comprehensive, in particular for elementary students whose writing often contains technical, creative, and development pieces.



The chart titled "Intrinsic Cognitive Load Comparison" visually compares the average scores given by Human and AI graders for the Content component of elementary-level English essays. Chart Summary:

- Human graders: 0.81 average score
- AI graders: 1.62 average score

Intrinsic cognitive load, when it applies to this case, is about the inherent complexity of the material, the intellectual demand to create ideas, think and express meaning in terms of essay content. Likewise, AI graders might be giving more credit to content versus human evaluators, as they are more lenient and rarely vary in the way they evaluate idea development, even in less formal writing. As the AI scores go up, this can also give learners positive reinforcement that the content quality is not so difficult. While this could also cause an underestimation of the complexity necessary in academic writing, this is also an important advantage of SPU. On the other hand, human scoring is a little strict, they would critically assess the depth, coherence and originality of the semantic quality of an essay, which can only be possible with a more grounded understanding of what it is. However, this approach provides more appropriate feedback, but it is only appropriate because it decreases the students' cognitive load. However, this disparity in scoring reminds us that AI may serve to lessen the psychological burden in feedback, as the feedback is more achievable, but educators need to be aware that additional successes can come with greater overconfidence or exaggerating the student's understanding of what constitutes quality.

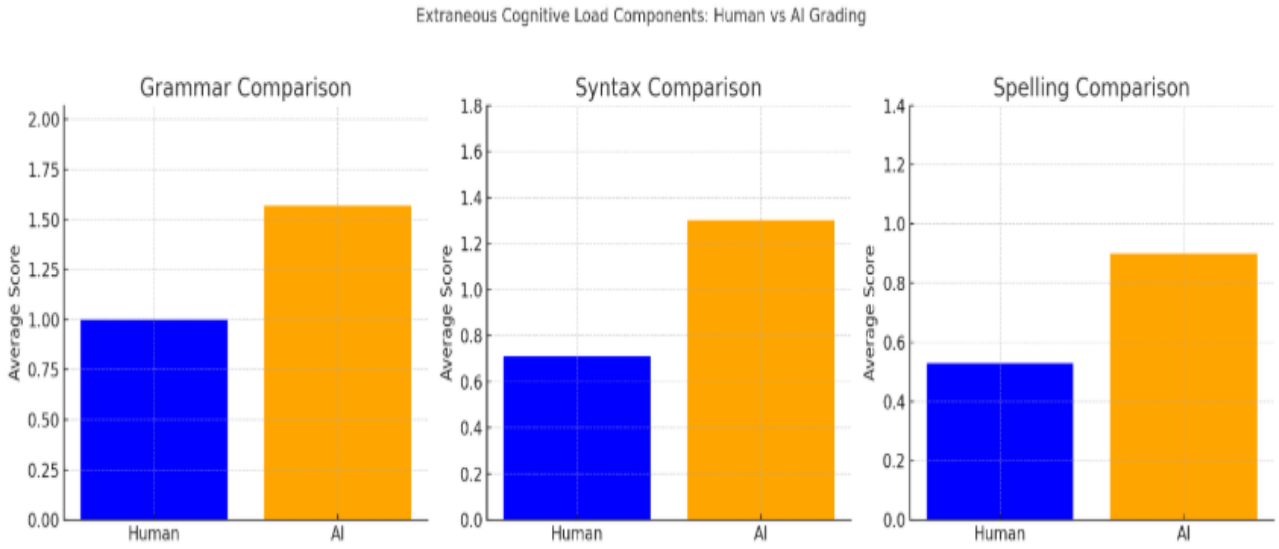



The chart titled "Extraneous Cognitive Load Comparison" presents the average scores assigned by Human and AI graders for components related to grammar, syntax, and spelling in student essays. These elements represent extraneous cognitive load, which refers to the mental effort required to process information that is not central to learning but still necessary, like formatting, correctness, and surface-level mechanics.

- Human Graders: 0.74 average score
- AI Graders: 1.27 average score

The differences between this chart tell us about the peripheral yet essential mechanics of writing, which each grading method handles differently.

This is because AI, being efficient in dealing with grammar, syntax, and spelling checks at face value, scores much higher. It takes away extraneous cognitive load for the students by giving immediate, automated, but surface-level feedback so that there's no more cognitive load for the students on doing things that might not necessarily be beneficial and take that away and devote that to higher-order things, such as idea development or idea refinement. Compared to human graders, human graders are more critical in a formative scenario and show more variability in judgment, which may lead to inconsistencies resulting in more cognitive load and anxiety due to correctness for the students. AI tools for integration can automate tasks such as grammar checking, freeing the teacher's time to consider more significant writing skills, however, educators need to stay engaged with students in language rules during AI-generated feedback. By reducing unnecessary workload on working memory, AI can provide more efficient feedback to learners than would otherwise be the case, creating a much more balanced cognitive environment for learning.




Here are the individual bar charts comparing Human and AI grading for each component of Extraneous Cognitive Load: Grammar, Syntax, and Spelling.

**Grammar**:

- AI average: 1.57
- Human average: 1.00

AI systems are more generous or consistent in grammar evaluation, possibly due to automated rule-based detection. This could help reduce extraneous load by streamlining feedback on grammatical errors.

**Syntax**:

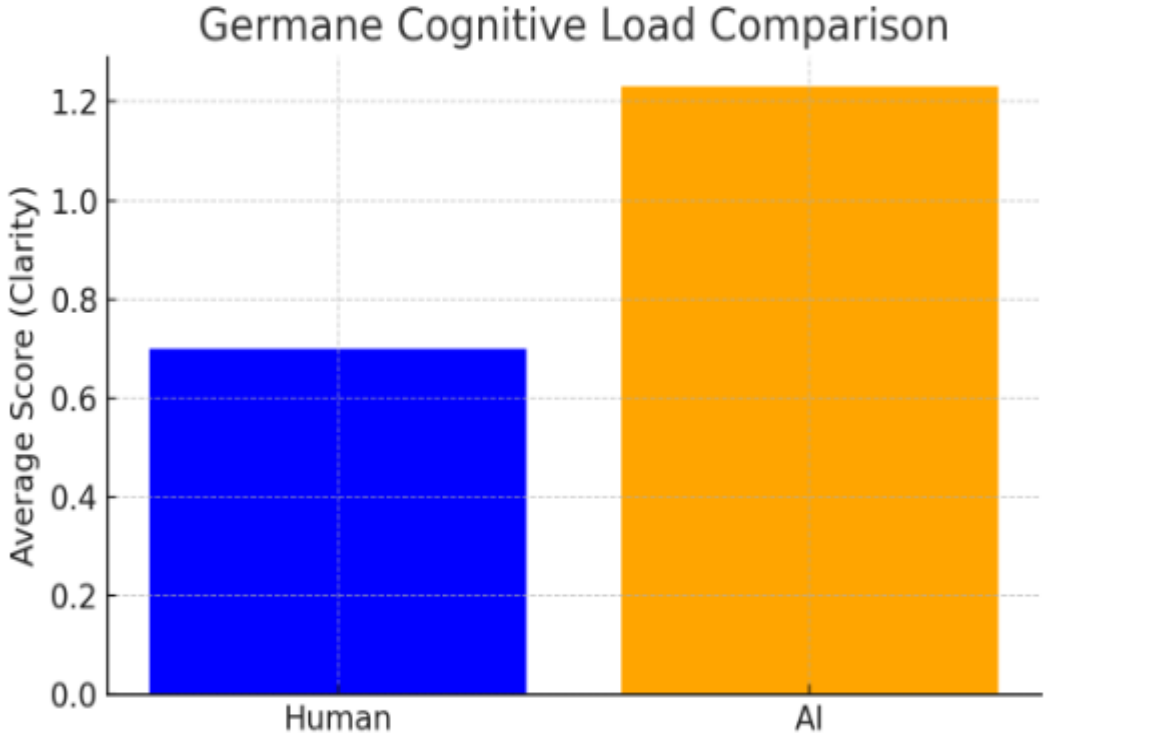
- AI average: 1.30
- Human average: 0.71

AI offers higher syntax scores, suggesting it may identify syntactic structure more favorably or with less bias. This can aid students in receiving quick, structured feedback without overloading their working memory.

**Spelling**:

- AI average: 0.90
- Human average: 0.53

AI excels in spotting and correcting spelling errors due to built-in dictionaries and language models. This significantly reduces extraneous load by automating a low-level but necessary task. These findings reinforce the idea that AI systems can effectively lower extraneous cognitive load by efficiently managing mechanical aspects of writing.




The chart titled "Germane Cognitive Load Comparison" illustrates the difference in average clarity scores given by Human and AI graders. Germane cognitive load refers to the mental effort invested in processing, understanding, and constructing meaningful learning. In writing, this includes the clarity of ideas, logical flow, and structural coherence.

- Human average score (Clarity): 0.70
- AI average score (Clarity): 1.23

This tendency for clarity in AI systems (whose scoring consistently rewards clarity, and organises, legible expression of pattern, sentence transitions and logical structure throughout the essay) is because of the systems' ability to pick out pattern-based coherence in what is written. However, human graders may be more conservative in grading clarity and consequently have fewer fitting procedures, inflicting variability in the scoring. AI feedback for learners could help them pay attention to ideation and the transmission of idea work, thus strengthening structured writing. While AI can be used to assist in concept shaping, it must be underpinned with the right to ensure that it doesn't take away clarity and become formulaic writing, and fails to recognise nuanced logical flaws. Supporting germane cognitive load has a significant role in supporting clarity, a key factor in the constructive use of understanding to support schema formation, a major goal of meaningful learning. This means that if AI feedback helps students to improve clarity, it may contribute to learning outcomes. Nevertheless, human intervention is still needed to pick up some of the extra subtleties of tone, argument strength and creative expression that AI may ignore.

**Integrative Link to Cognitive Load**

By managing efficiently and correcting surface-level issues like grammar, syntax and spelling, AI consistently outperformed human graders on the Extraneous Load Comparison chart (grammar, syntax, spelling), lowering the cognitive load on the students. So students are freed up to shift working memory and attention to more germane tasks like clarifying message intent, organising arguments, and lifting higher AI clarity scores regarding coherence elements of correct outputs. Like the Intrinsic Load Comparison, the AI was better at the content of a structure than its perfection, showing that AI finds usable ideas even if the structure is not perfect. In this approach, the students tend to feel more confident exploring and expressing their thoughts, thereby reducing the scary aspects of an intrinsically complex topic. Focusing on germane load by tackling both the extraneous and intrinsic load in a more supportive way supports students in refining clarity and meaning. Taken together, these findings imply that the use of AI grading can be part of a scaffolded environment that allows the removal of extraneous load, the adjustment of intrinsic load, and the optimisation of germane load; thus, uniquely, AI can be useful for educators to promote more meaningful and deeper learning through writing.

**Reliability Analysis**

Cronbach's Alpha measures the internal consistency or reliability of a set of items (in this case, grading criteria) and helps determine how consistently the evaluators apply the scoring rubric.

**Cronbach's Alpha Results**

- Human Evaluations: The Cronbach's Alpha for human evaluations across the five criteria is 0.83, indicating a high level of internal consistency and reliability in how the human evaluators apply the scoring rubric.
- AI Evaluations: The Cronbach's Alpha for AI evaluations is 0.78, which also reflects a moderate to high level of internal consistency. However, the AI's consistency is slightly lower than that of the human evaluators.

**Interpretation:**

- A Cronbach's Alpha above 0.7 is generally considered acceptable, with values closer to 1 indicating excellent reliability.

- Both the Human and AI systems show good reliability in their grading of the essays, with Humans being slightly more consistent in their assessments.

This reliability analysis suggests that both evaluators (Human and AI) are relatively consistent in their application of the grading rubric, with Human evaluators showing a marginally higher consistency.

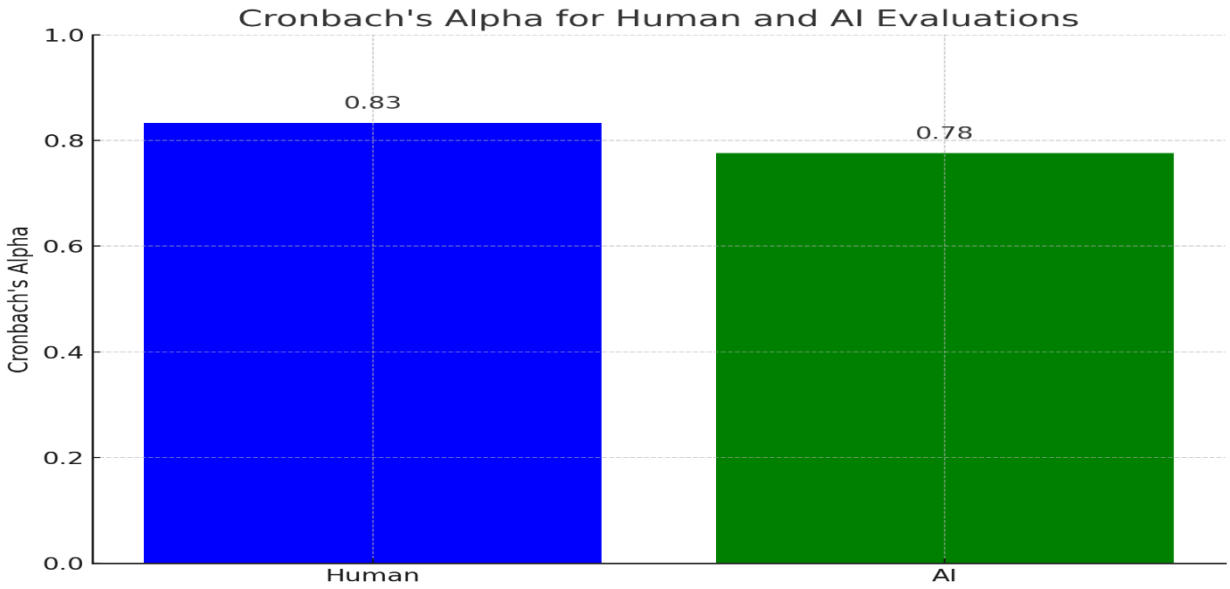



Figure 5: Cronbach's Alpha for human and AI evaluations

**Qualitative Analysis**

Cognitive Load Theory states that human working memory has a limited capacity, and that cognitive load should be managed adequately to avoid overflow in the optimisation of learning. CLT describes three types of cognitive load: intrinsic load related to the difficulty of the task, extraneous load related to way the information is presented, and germane load (in relation with learning effort and building schema).

The process of evaluating an essay puts human evaluators in a cognitive load management problem. They are not only required to assess grammar and syntax, but also what is more subjective of clarity and content relevance. Therefore, human evaluators must evaluate a whole essay, which increases their cognitive load. For example, to test clarity, you need to understand the structure, style and message communicated in the text, which requires a greater cognitive load than the mechanical evaluations technique, such as grammar.

Like CLT and constructivism, this idea suggests that learning and assessment are not simply about straightforward evaluation. When it comes to higher cognitive demands, humans can tackle personal interpretations, such as issues of clarity and style. However, human scoring can introduce extraneous cognitive load, resulting in inconsistencies. Variations in human judgment can arise from the different mental schemas of evaluators, as well as the time and attention needed to discern the subtle points and ambiguities in each student's work.

On the other hand, AI systems are built to perform tasks based on a set of defined rules, namely grammar and syntax, so the cognitive load is pushed to a minimum. Because these systems can assess objectively aspects of the essay so quickly and efficiently, they can quickly and efficiently

give high scores for grammar and syntax without much cognitive effort. That is what makes AI so powerful of being able to crunch huge amounts of data, providing consistent and objective feedback rather than needing to create complicated interpretations.

Nevertheless, there are limitations on how AI works with more subjective criteria such as clarity and content. Therefore, AI cannot directly reproduce that well of deep comprehension and contextual understanding that human evaluators have over the text. For instance, it helps with spotting trends and doing lexical analysis, but not with the sense of creativity, intent and originality as humans. In a sense, then, AI is not equipped or able to work with the germane cognitive load that humans themselves bring to bear through their creation of schemas, which inform creativity in context.

## Discussion

The patterns of grading from AI and human evaluators were compared for each of the key linguistic features, grammar, syntax, content, and clarity, amongst themselves, focusing on how they handled grammar, syntax, content, and clarity. This analysis examines these features in the light of Cognitive Load Theory (CLT) and provides significant differences depending on how each type of evaluator processes the same features. This is the reason why AI systems do well when it comes to grammar, syntax, and even spelling, because these are what you call low cognitive load. There is also a clear rule and pattern that governs these aspects and can be encoded in the algorithm for the speedy and accurate processing and assessment by AI. Such tasks are designed with a low cognitive load, which is in line with CLT's notion that intrinsically simpler tasks (requiring fewer cognitive resources to interpret and reason) are easier to automate. As a result, AI grading trends are very consistent and very objective when judging the structure of student writing.

Human evaluators are more effective for higher cognitive load tasks like content and clarity. The subject is based more on the deeper cognitive process involved, it requires a heavier understanding of the context, creativity and the essay's message. The second view is a resting point in the CLT debate, the situation in which the task is too high in cognitive load (evaluating a person's creativity or the coherence of an argument) and requires extracting personal experience, prior knowledge and contextual understanding, which human evaluators can do. It is here that human evaluators' strengths lie in that they can interpret nuances in the writing, for example, argument depth and flow of ideas and originality that are difficult to spot by machines. CLT's description of these tasks suggests a lot of such expensive, inherently non-rule-based cognitive resources that are not easily doable in a machine setting. This distinction indicates that AI and human evaluators treat high cognitive load (creative) and low cognitive load (structured) parts of writing quite differently.

Moreover, the analysis explores what the hybrid system consisting of AI and evaluators should look like and assesses its feasibility and potential benefits. Because AI systems can evaluate mechanical aspects very quickly, they can help greatly reduce extraneous cognitive load by automating the evaluation of grammar, syntax and spelling. It frees up the human evaluators to use their cognitive resources in deeper understanding, critical thinking, and better comprehension of the architecture and the arguments as a whole. Through the use of a hybrid system, AI's efficiency can be combined with the interpretative abilities of human evaluators to optimise cognitive load management and improve consistency, as well as to improve the entire grading process. As a means of examining how students perform during examinations, compared grading pattern analyses, cognitive load analysis, and investigation of possible advantages of a hybrid grading system in the academic sphere, this approach is in line with the study's goals.

## Conclusion

This study compared grading patterns of AI and human evaluators across key linguistic features and was also used to analyse cognitive load for each under structured and creative aspects of student writing. The findings indicate that AI is very strong in assessing mechanical aspects of grammar, syntax, and spelling, which are low cognitive load tasks. This facility, led by AI to apply known criteria and algorithms, means that quick, objective, consistent feedback can be given that takes the cognitive load of both the author and user. However, with subjective criteria such as clarity, depth of content, and creativity, what AI can output is limited to a point because the judgments need to be made on a human level and therefore require nuanced interpretation and contextual understanding that humans can provide.

It also mentions the opportunities of a hybrid grading system, in which AI can help with grading recognisable, lower cognitive load items, and human assessors work with more subjective, high cognitive load issues. This would optimise cognitive load management, hence optimising the grading process more efficiently and proportionately. Such a system, with AI's efficiency coupled with human evaluators' capacity to rate creativity and content, would likely improve grading accuracy, deal with the inconsistencies, and make the overall feedback process more meaningful in terms of education.

## References

1. Aghaziarati, A. (2023). Artificial intelligence in education: Investigating teacher attitudes. aitechbesosci, 1(1), 35-42. https://doi.org/10.61838/kman.aitech.1.1.6
2. Alharbi, W. (2023). AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools. Education Research International, 2023, 1-15. https://doi.org/10.1155/2023/4253331
3. Alshehri, B. (2023). Pedagogical paradigms in the AI era: insights from Saudi educators on the long-term implications of AI integration in classroom teaching. IJESA, 2(8), 159-180. https://doi.org/10.59992/ijesa.2023.v2n8p7
4. Attali, Y. (2014). Reliability-based feature weighting for automated essay scoring. Applied Psychological Measurement, 39(4), 303-313. https://doi.org/10.1177/0146621614561630
5. Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).
6. Bhowmik, S. (2024). Teaching elementary ESL writing in Canada. TESL Canada Journal, 40(1), 107-135. https://doi.org/10.18806/tesl.v40i1/1387
7. Chen, C., & Gong, Y. (2025). The Role of AI-Assisted Learning in Academic Writing: A Mixed-Methods Study on Chinese as a Second Language Students. *Education Sciences*, *15*(2), 141.
8. Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: a multivariate generalizability theory application. American Educational Research Journal, 59(6), 1122-1156. https://doi.org/10.3102/00028312221106773
9. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.
10. Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International journal of English studies*, *1*(2), 1-23.
11. Cummins, R., Zhang, M., & Briscoe, T. (2016). Constrained multi-task learning for automated essay scoring. https://doi.org/10.18653/v1/p16-1075

12. Dergaa, I., Chamari, K., Żmijewski, P., & Saad, H. (2023). From human writing to artificial intelligence-generated text: Examining the prospects and potential threats of Chatgpt in academic writing. Biology of Sport, 40(2), 615-622. https://doi.org/10.5114/biolsport.2023.125623
13. Familoni, B. (2024). Advancements and challenges in AI integration for technical literacy: a systematic review. Engineering Science & Technology Journal, 5(4), 1415-1430. https://doi.org/10.51594/estj.v5i4.1042
14. Feng, L. (2025). Investigating the effects of artificial intelligence-assisted language learning strategies on cognitive load and learning outcomes: A comparative study. *Journal of Educational Computing Research*, *62*(8), 1961-1994.
15. Firdaus, R., Xue, Y., Gang, L., & Sibt e Ali, M. (2022). Artificial intelligence and human psychology in online transaction fraud. *Frontiers in psychology*, *13*, 947234.
16. Ge, H., Wu, P., Dong, L., OuYang, N., Chen, J., & Chen, J. (2025). Takeover performance prediction model considering cognitive load: analysis of subjective and objective factors. *Traffic Injury Prevention*, 1-9.
17. Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy. *Brain Sciences*, *15*(2), 203.
18. Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S., … & Koedinger, K. (2021). Ethics of AI in education: towards a community-wide framework. International Journal of Artificial Intelligence in Education, 32(3), 504-526. https://doi.org/10.1007/s40593-021-00239-1
19. Huang, R. (2024). Exploring AI tools in early childhood education: usage patterns, functions, and developmental outcomes. https://doi.org/10.5772/intechopen.1007116
20. Hussein, M., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: a literature review. Peerj Computer Science, 5, e208. https://doi.org/10.7717/peerj-cs.208
21. Hussein, M., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: a literature review. Peerj Computer Science, 5, e208. https://doi.org/10.7717/peerj-cs.208
22. IELTS. (2023). *IELTS writing band descriptors*. Retrieved from https://www.ielts.org
23. Joo, S. (2024). Generative AI as writing or speaking partners in L2 learning: implications for learning-oriented assessments. Studies in Applied Linguistics and TESOL, 24(1). https://doi.org/10.52214/salt.v24i1.12865
24. Mane, P. C. (2025). Accuracy and creativity analysis of Chatgpt in quantitative aptitude. *The International Journal of Information and Learning Technology*.
25. Matsumura, L., Correnti, R., & Wang, E. (2015). Classroom writing tasks and students' analytic text-based writing. Reading Research Quarterly, 50(4), 417-438. https://doi.org/10.1002/rrq.110
26. Mohammadkarimi, E., Omar, J. A., & Rashid, A. S. (2025). Advancements and Challenges in Automated Evaluation of Spoken Language Proficiency Using AI. In *Using AI Tools in Text Analysis, Simplification, Classification, and Synthesis* (pp. 229-260). IGI Global Scientific Publishing.
27. Mohammed, A. (2023). Examining the implementation of artificial intelligence in early childhood education settings in Ghana: educators' attitudes and perceptions towards its long-term viability. American Journal of Education and Technology, 2(4), 36-49. https://doi.org/10.54536/ajet.v2i4.2201
28. Nadeem, F., Nguyen, H., Liu, Y., & Ostendorf, M. (2019). Automated essay scoring with discourse-aware neural models. https://doi.org/10.18653/v1/w19-4450
29. National Writing Project (NWP). (2019). *The NWP rubric for assessing writing*. Retrieved from https://www.nwp.org

30. Ozer, O. (2025). Understanding Translation Students' Use of AI-Powered Tools During Text Revision and Their Impact on Cognitive Load. In *Reimagining Intelligent Computer-Assisted Language Education* (pp. 281-308). IGI Global.
31. Park, W. and Kwon, H. (2023). Implementing artificial intelligence education for middle school technology education in the Republic of Korea. International Journal of Technology and Design Education, 34(1), 109-135. https://doi.org/10.1007/s10798-023-09812-2
32. Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: Chatgpt and beyond. Journal of University Teaching and Learning Practice, 20(2). https://doi.org/10.53761/1.20.02.07
33. Poupard, M., Larrue, F., Sauzéon, H., & Tricot, A. (2025). A systematic review of immersive technologies for education: effects of cognitive load and curiosity state on learning performance. *British Journal of Educational Technology*, *56*(1), 5-41.
34. Ramesh, D. and Sanampudi, S. (2021). An automated essay scoring system: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527. https://doi.org/10.1007/s10462-021-10068-2
35. Shermis, M. D., & Burstein, J. (2013). Handbook of automated essay evaluation. *NY: Routledge*.
36. Smedt, F., Merchie, E., Barendse, M., Rosseel, Y., Naeghel, J., & Keer, H. (2017). Cognitive and motivational challenges in writing: studying the relation with writing performance across students' gender and achievement level. Reading Research Quarterly, 53(2), 249-272. https://doi.org/10.1002/rrq.193
37. Sweller, J. (1988). Cognitive load during problem-solving: Effects on learning. *Cognitive science*, *12*(2), 257-285.
38. Sweller, J., Van Merrienboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*, 261-292.
39. Tobing, V. S. B. L., & Damanik, B. A. R. (2025). JOURNAL ENTRY: THE IMPACT OF COGNITIVE LOAD ON REAL-TIME LANGUAGE PRODUCTION. *Sabda: Jurnal Sastra dan Bahasa*, *4*(1), 1-7.
40. Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Vol. 86). Harvard University Press.
41. Wale, B. (2024). The transformative power of AI writing technologies: enhancing EFL writing instruction through the integrative use of Writerly and Google Docs. Human Behaviour and Emerging Technologies, 2024, 1-15. https://doi.org/10.1155/2024/9221377
42. Wang, C. (2024). Exploring students' generative AI-assisted writing processes: perceptions and experiences from native and non-native English speakers. Technology, Knowledge and Learning. https://doi.org/10.1007/s10758-024-09744-3
43. Wang, E., Matsumura, L., & Correnti, R. (2017). Written feedback to support students' higher-level thinking about texts in writing. The Reading Teacher, 71(1), 101-107. https://doi.org/10.1002/trtr.1584
44. Wong, W. and Bong, C. (2019). A study for the development of automated essay scoring (AES) in the Malaysian English test environment. International Journal of Innovative Computing, 9(1). https://doi.org/10.11113/ijic.v9n1.220
45. Wu, Y., Henriksson, A., Nouri, J., Duneld, M., & Li, X. (2022). Beyond benchmarks: spotting key topical sentences while improving automated essay scoring performance with topic-aware Bert. Electronics, 12(1), 150. https://doi.org/10.3390/electronics12010150
46. Xie, W., Li, J., Mu, Y., Zhang, H., Zhao, S., Zheng, X., & Song, K. (2025). The Power of Personalised Datasets: Advancing Chinese Composition Writing for Elementary School through Targeted Model Fine-Tuning. *International Journal of Asian Language Processing*, 2450017.

47. Yu, L. and Yu, Z. (2023). Qualitative and quantitative analyses of artificial intelligence ethics in education using vosviewer and citnetexplorer. Frontiers in Psychology, 14.