Volume 7 Issue 27 (October 2025) PP. 190-206
DOI: 10.35631/IJMOE.727013

INTERNATIONAL JOURNAL OF
MODERN EDUCATION
(IJMOE)
www.ijmoe.com

THE EFFECTIVENESS OF AI IN ASSESSING ESL UNIVERSITY STUDENTS' ARGUMENTATIVE WRITING IN COMPARISON TO RUBRIC-BASED ASSESSMENT AND TOULMIN'S MODEL OF ARGUMENTATION

Farous Izwan Abdul Aziz[1*], Seriaznita Mat Said[2]

[1] Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah Kampus Antarabangsa Labuan (UMSKAL), Malaysia,
Email: farousizwan.aziz@ums.edu.my
[2] Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia (UTM), Kuala Lumpur, Malaysia,
Email: seriaznita.kl@utm.my
[*] Corresponding Author

Article Info:

Article history:
Received date: 26.06.2025
Revised date: 10.07.2025
Accepted date: 25.08.2025
Published date: 01.10.2025

To cite this document:

Abdul Aziz, F. I., & Mat Said, S. (2025). The Effectiveness of AI in Assessing ESL University Students' Argumentative Writing in Comparison to Rubric-Based Assessment and Toulmin's Model of Argumentation. *International Journal of Modern Education, 7* (27), 190-206.

DOI: 10.35631/IJMOE.727013

Abstract:

Mastering argumentative writing is essential for ESL university students, yet many continue to face challenges in structuring logical arguments, providing evidence, and articulating ideas effectively. This study examines their writing proficiency through rubric-based assessment and Toulmin's Model of Argumentation, while also comparing AI-assisted grading with human evaluation. A total of 72 student essays were analysed to assess argument structure, reasoning, and language proficiency. Results indicate that while students consistently present claims and grounds, they often struggle with incorporating warrants, backing, and rebuttals. Additional issues include weak organisation, repetitive reasoning, and frequent grammatical errors that impair clarity. AI-based grading demonstrated efficiency in evaluating grammar and structure but tended to be stricter than human evaluators. The findings highlight the importance of explicit instruction in argumentation, the integration of Toulmin's Model into ESL pedagogy, and targeted grammar interventions. AI assessment tools, while operationally beneficial, should augment rather than supplant human evaluators to maintain holistic judgment. Strengthening these areas can enhance students' critical thinking and overall writing performance.

Keywords:

ESL Writing, Argumentative Writing, Toulmin's Model, AI Assessment, Rubric-Based Evaluation

Introduction

Argumentative writing skills are vital for university students, especially for those enrolled in English as a Second Language (ESL) courses. A compelling argument relies on logical organisation, critical thinking and strong language proficiency. Unfortunately, many ESL students struggle in these areas despite over a decade of formal instruction (Ramli, Abdul Kadar, & Rafek, 2024). The Malaysian Ministry of Higher Education (MOHE) emphasises the importance of mastering argumentative writing (Malaysian Education Blueprint, 2015). Despite this, students still face challenges in structuring their arguments, providing evidence and expressing their ideas clearly (Sabri et al., 2021; Joannes & AlSaqqaf, 2022). This is especially true for second language (L2) students who still struggle to master English proficiency (Campbell & Filimon, 2018; Pei, Zheng, Zhang, & Liu, 2017) as early as the primary level whereby L2 students are reported to experience difficulty in mastering vocabulary, inability to spell words correctly, and first language (L1) interference (Ghulamuddin, Mohari & Ariffin, 2021). This deficiency underscores the need for a more comprehensive assessment of ESL students' argumentative writing proficiency. This study critically compares the reliability of AI assessment and human evaluation of argumentative writing skills of degree-level university students by employing rubric-based assessment and Toulmin's Model for Argumentation (Toulmin, 2003). Additionally, the study compares AI-based grading with human evaluation to assess the reliability of automated writing assessment tools.

Argumentative writing is a challenging genre for ESL students as they are required to confirm their stance, support their claims with logical reasoning, and engage in counterarguments (Aziz, Salam & Mat Said, 2023; Aziz, 2021; Aziz & Mat Said, 2020a, 2020b, 2019; Aziz & Ahmad, 2017; Sabri et al., 2021). Despite being exposed to English writing during their secondary school years, many students enrolled in university still lack the skills to develop strong, evidence-based arguments (Bahari et al., 2021; Ramli et al., 2024). This outcome is a result of schools being exam-oriented, which teaches students how to answer exams rather than developing critical thinking skills (Aziz & Mat Said, 2019). Therefore, students need to cultivate critical thinking skills as university-level courses require them to compose well-reasoned academic reports with strong arguments supported by credible evidence (Sundari & Febriyanti, 2021; Goldman, 2019).

Toulmin's Model (2003) is widely used to improve argumentative writing instruction as it provides a straightforward approach to composing logical arguments (Aziz, Salam, & Mat Said, 2023). Additionally, rubric-based assessment also offers clear evaluation criteria, which allow students to identify key areas for improvement (Rock, 2021). With the advancement of Artificial Intelligence (AI), automated writing assessment tools have been introduced to provide accelerated feedback and complement traditional grading methods (Almegren et al., 2024). This, however, raises concerns about AI's ability to accurately evaluate the thematic consistency and quality of the students' arguments compared to human evaluators (Jonäll, 2024).

This study aims to (1) evaluate ESL students' persuasive writing proficiency using a rubric-based assessment, (2) analyse the structural effectiveness of arguments using Toulmin's Model, and (3) compare AI-based grading with human evaluation in identifying key weaknesses in students' argumentative writing.

This study offers a look into the argumentative writing challenges faced by ESL students and provides recommendations to improve their argumentation skills, organisation, and language proficiency. In addition, this study contributes to the ongoing discussion on the role AI plays in academic writing assessment, determining its potential to augment human evaluation (Bouziane & Bouziane, 2024). As outlined by UNESCO (2019, 2021, 2023), human oversight remains essential in AI-augmented assessment to mitigate algorithmic bias. By highlighting students' weaknesses and evaluating the effectiveness of traditional and AI-assisted assessment methods, this study aims to improve ESL writing instruction and foster the development of critical thinking in university students.

Literature Review

Argumentative writing, which is an essential skill that students need to succeed and survive in today's world (Sabri et al., 2021; Joannes & AlSaqqaf, 2022), is one of the many genres included in the English second language (ESL) syllabus (Bahari, Kussin, Harun, Mohamed & Jobar, 2021). However, argumentative writing is a challenging genre to master (Sabri et al., 2021). This complexity arises from the dual cognitive demands of critical issue analysis and opinion formation, requiring students to engage substantively with opposing viewpoints (Joannes & AlSaqqaf, 2022). The Ministry of Higher Education (MOHE) places great emphasis on students mastering argumentative skills; however, students are still experiencing difficulties (Sabri et al., 2021). This is because the Malaysian Education Blueprint 2013-2025 (Ministry of Education (MOE), 2013) primarily focuses on developing students' proficiency in their second language (L2).

Just as they did in secondary school, students at the tertiary level are faced with multiple genres of academic writing, including discursive, problem-solving, and argumentative essays (Ramli, Abdul Kadar & Rafek, 2024). However, despite having experienced ESL training since their primary school years, which should have made them well-equipped, they are still hindered by their limited English vocabulary, poor grammatical skills and difficulties in generating ideas (ibid). Various internal and external factors, such as the complexity of argumentative writing, inconsistent assessment criteria, and limited engagement in practical writing activities, further compound these challenges. Students' writing proficiency is also influenced by the frequency of writing practice, support from teachers, their general knowledge and reading habits, all of which are essential for producing relevant content for their arguments. Additionally, language-based games and activities have been shown to positively impact their persuasive writing development (Aziz, 2021).

Over the years, numerous models and strategies have been developed as solutions to address argumentative writing challenges and enhance argumentative writing skills (Flower & Hayes, 1981; Fei-Wen, 2010; Harris & Graham, 2018; Sampson et al., 2013; Manz, 2015). However, Toulmin's Model (Toulmin, 2003; Hitchcock, 2017) is widely considered the most definitive framework for argumentative writing due to its practicality and potential to enable students to structure arguments and generate ideas effectively (Aziz, Salam & Mat Said, 2023; Aziz & Mat Said, 2020; Hitchcock, 2017).

According to Toulmin (2003), an argument is composed of two sets of elements. The first set of elements includes claim, grounds and warrants, which provide the foundation of the argument. A claim is the main argument or thesis, while the grounds are evidence or reasons that support the claim. Finally, the warrant is the logical connection between the claim and the grounds. The second set of elements helps strengthen the foundation. These comprise backing, which provides additional support for the warrant; qualifiers, which are words or phrases that indicate the strength of the claim being asserted (e.g., "probably", "likely"), and finally, rebuttals, which are counterarguments or limitations to the claim, also known as opposing perspectives. As a tool for content analysis, Toulmin's Model is capable of discerning the deficiencies within written arguments. Table 1 shows the list of research studies related to Toulmin's Model.

Table 1: Past Studies in Toulmin's Model in Argumentation

Author(s) & Year	Objective	Methodology	Key Findings
Marsaulina, 2024	Assess essay writing using Toulmin's model	47 first-year engineering students, content analysis	Weak essays: missing/unclear claims, weak evidence, no rebuttals
Indarti, 2025	Compare Toulmin use by gender & topic familiarity	ICNALE corpus, descriptive analysis	Females: stronger arguments; Males: basic only; Chinese: balanced; Familiar topics = better quality
Ghoneam & Badawy, 2023	Test Toulmin's model in debates	80 EFL students, Egypt; pre/post test, 10-week intervention	Improved structure, richer reasoning, stronger argumentation in the experimental group
Osman & Januin, 2021	Analyse argumentative essays	18 essays, 15 ESL undergrads, Malaysia	Rebuttals absent
Novianti, 2021	Examine response essays	29 undergrads, purposive sample of 16 essays	Few used warrants (<7); very few rebuttals (3); weak analytical skills

Rubric-based writing assessment is widely recognised as a tool for evaluating written works both reliably and accurately (Toscu, 2023). Across the world, the assessment of English writing skills is considered a fundamental aspect of ESL education, with grammar, vocabulary, sentence structure, coherence, and clarity being key elements (Zahid, Anwar & Azim, 2023). Rubrics play a critical role in enhancing student writing skills by fostering awareness of writing criteria, promoting self-regulation, encouraging critical thinking, and facilitating peer feedback that supports writing skills through structured guidance and scaffolding (Kunnel, 2021).

Furthermore, rubric-based assessment provides clear guidance for ESL learners, facilitating comprehension, self-awareness, and self-improvement via self-assessment. However, its effectiveness depends on the students' engagement and reinforcement from the teacher's feedback (Dyrdal, 2021). In addition to assessment, rubrics also serve as a reference tool for language learners, strengthening their genre knowledge and tracking their progress in genrespecific writing (Rock, 2021). More importantly, rubrics constructed based on actual performance data can provide a valid method for assessing second language writing and function as a learning tool that guides students in developing their writing skills (Rock, 2021).

Manually grading exam papers and written assignments is a crucial yet time-consuming and tedious task for educators, often leading to inconsistencies due to evaluator fatigue (Seßler et al., 2025; Vijaya Shetty et al., 2022). Artificial Intelligence (AI) has the potential to provide high-quality and consistent feedback, free from human biases, while also assessing a large volume of essays within a short period (Almegren et al., 2024). However, despite its efficiency, AI is unable to recognise the nuanced aspects of students' writing that human evaluators can discern (Jonäll, 2024). In a similar vein, the research by Wetzler et al. (2024) further highlights discrepancies between grades assigned by human instructors and those assigned by AI, even when using the same rubric. Consequently, UNESCO, during the Beijing Consensus on Artificial Intelligence and Education (2019), in its effort to support the application or development of AI tools, stresses the need "to support adaptive learning processes; to leverage the potential of data to enable the evaluation of the multiple dimensions of students' competencies; and to support large-scale and remote assessment." (pp. 5-6). UNESCO (2023) welcomes the application of AI-based assessment tools in higher education learning, particularly in a positive light. AI-assisted systems use a generative pre-trained transformer, "a type of large language model (LLM) that is pre-trained on even larger amounts of data, which allows the model to capture the nuances of language and generate coherent context-aware text." (UNESCO, 2023, p. 8) AI-based assessment tools demonstrate superior performance in evaluating fundamental language mechanics, including grammar, spelling, sentence structure, relevance, and supporting evidence. Although AI-based programs offer speed and efficiency in assessment, they are not without limitations. Human evaluators continue to outperform AI in assessing thematic consistency, which remains a critical aspect of comprehensive evaluation (Bouziane & Bouziane, 2024). Hence, in educational practice, AI ought to serve in a supplementary capacity—enhancing assessment efficiency while preserving educators' irreplaceable role in evaluative discernment. Table 2 shows the list of related research studies.

Table 2: Past Studies Related to AI vs. Manual Essay GradinG

Author(s) & Year	Objective	Methodology	Key Findings
Uyar & Büyükahıska, 2024	Compare AI vs human scoring	EFL essays, Turkey; e-rater vs human raters	High agreement overall; AI weaker in argument depth & cohesion
Quah et al., 2024	Test ChatGPT (GPT-4) reliability	69 dental students, Singapore; AES vs human scores	Reliable for one essay, moderate for another; AI close to humans in one, lower in the other
Seßler et al., 2025	Compare LLMs vs teacher ratings	20 essays, Germany; 5 LLMs vs 37 teachers	Closed-source LLMs aligned better; strong in language, weaker in content assessment

Research Methodology

This study employed a mixed-methods research design. The study took place in a local university in the Federal Territory of Labuan, Malaysia. Seventy-two essay samples were collected from degree-level university students during the final week of a semester-long (14 weeks) Reading and Writing English course. Before the students wrote their essays, they received weekly instruction on how to write various essay genres, with a particular emphasis on argumentative writing.

During each class, the students were exposed to the elements of Toulmin's Model and how they functioned together. First, they were introduced to the elements, and then they were gradually taught how to employ each of these elements to construct their arguments. By the end of the semester, students were instructed to write essays based on a set of prompts provided by the instructor. Mobile devices were collected to prevent plagiarism.

The process is illustrated in the flowchart in Figure 2.

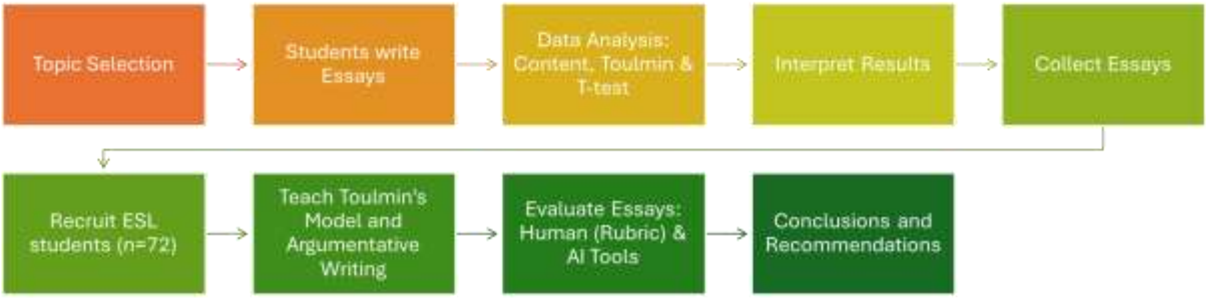
Topic Selection
Students write Essays
Data Analysis: Content, Toulmin & T-test
Interpret Results
Collect Essays
Recruit ESL students (n=72)
Teach Toulmin's Model and Argumentative Writing
Evaluate Essays: Human (Rubric) & AI Tools
Conclusions and Recommendations

Figure 1: Study Methodology Flowchart

This study involved a total of 72 ESL undergraduate students enrolled in a semester-long Reading and Writing English course at a Malaysian public university. The students were primarily from the Faculty of International Finance (FKAL), with a few from the Faculty of Computer and Informatics (FKI) and one from the Faculty of Islamic Studies (FIS). A convenience sampling method was employed, selecting participants based on their availability and enrollment in the relevant course. This non-probability sampling technique was chosen due to practical constraints, allowing access to students who had been exposed to Toulmin's Model and had received instruction on argumentative writing. Demographic information was collected from the students using a set of questionnaires. Table 3 depicts the demographic information that was collected from the students:

Table 3: Demographic Information of the Student Respondents (N=72)

Age		Gender		Faculty		Language Spoken at Home		MUET Scores	
20	10	Male	57	FKI	13	English	0	2	1
21	54	Female	14	FKAL	58	Malay	56	2.5	1
22	3	Not Mentioned	1	FIS	1	Chinese	6	3	19
23	4					Tamil	5	3.5	50
Not Mentioned	1					Others	4	Not Mentioned	1

The table shows that the students' ages range between 20 and 23, with most of them being 21 years old. However, one student failed to mention their age. The table also lists the genders of the students, with a majority of them being female. However, one student failed to mention their gender. The faculties to which the student respondents belong are also listed. A majority of them enrolled in the Labuan Faculty of International Finance (FKAL). One student, however, was not part of the Faculty of Computer and Informatics (FKI) or FKAL, but instead belonged to the Faculty of Islamic Studies (FIS). The table also depicts the languages spoken at home by the students, indicating that none speak English. A majority of them do speak Malay, with Chinese, Tamil and Other languages being the minority. Finally, a majority of the students have earned a score of 3.5 for their MUET scores, which falls between the modest (Band 3.0) and competent user (Band 4.0) levels.

Several data distribution challenges were observed during the study. First, there was a demographic imbalance, as a significant majority of participants were female (79.2%), which may have affected the gender representativeness of the findings. Additionally, over 80% of the students were from a single faculty (FKAL), limiting cross-disciplinary diversity. Another issue was the linguistic background, as none of the participants reported using English as their primary home language, which may have influenced their grammatical accuracy and fluency. In terms of proficiency level, most students had MUET scores between Band 3.0 and 3.5, representing intermediate English proficiency, with higher or lower bands underrepresented. Prompt selection bias was also noted, with nearly 47.2% of the students choosing the same essay topic, "Are online classes better than face-to-face classes?", which reduced variation in content and limited comparative depth across different argumentative contexts. Finally, a small number of participants did not report their age, gender, or language proficiency scores, slightly affecting the completeness of demographic analysis.

A rubric designed explicitly for assessing critical thinking in English argumentative essays was adopted for this study. Beyond serving as a reliable tool for evaluating second-language writing (Rock, 2021), the rubric also facilitated the identification of weaknesses in the students' essays based on clearly defined argumentative writing criteria (Kunnel, 2021).

The essays were assessed based on three main criteria. Organisation, allocated 12 marks, evaluated the quality of communication, logical reasoning, clarity of arguments, and the strength of conclusions. Content, worth 20 marks, focused on the relevance, depth, organisation, and maturity of ideas presented. Finally, Language and Vocabulary, carrying 18 marks, assessed grammar, sentence structure, vocabulary variety, and accuracy in word choice.

The evaluator, a university ESL instructor with 11 years of teaching experience, graded the essays. The full points that can be assigned to an essay are 50.

Along with the rubric, an AI-based assessment tool was used to grade the essays. The rubric was fed into the tool before each essay was uploaded for grading. An AI-based assessment program was also used to assist in analysing the essays. AI-based tools were selected due to their speed and efficiency in evaluating the large volume of written content.

Toulmin's Model (2003) was employed to examine the essays and verify how the students utilise the six persuasive elements: claim, grounds, warrant, qualifier, backing and rebuttal. Toulmin's Model analysis was aided by the use of an AI-based assessment tool that was coded for analysing argumentative essays.

Content Analysis was employed to examine the common features within the students' persuasive essays. This was facilitated by Toulmin's Model as the elements within each of the sample essays were identified, tagged and coded to examine how frequently the students utilise the elements from Toulmin's Model.

For this study, the t-test was used to compare the mean scores provided by the rubric-based human assessment and the AI assessment. This is to determine if there are any significant differences between the scores. Not only are the total mean scores compared, but the mean scores for "Organisation", "Content" and "Language & Vocabulary" were also compared.

Essays were collected from tertiary ESL students. The essays were written based on a set of prompts provided by the researcher. The prompts are listed below:

a. Do you agree that online classes are better than face-to-face classes? Support your opinion.
b. Should all COVID-19 restrictions be lifted? Support your opinion.
c. Should all school examinations be abolished? Support your opinion.
d. Should university course textbooks be provided for free? Support your opinion.
e. Are e-books better than physical textbooks? Support your opinion.

A rubric designed with critical thinking in mind was utilised to assess the essays, with AI-assisted analysis used to accelerate data categorisation and reduce personal bias in evaluation. The evaluator, using the rubric, ensured the reliability of the data. The rubric was not only used to assess the essays but also to reveal key areas of weakness within them based on the rubric criteria.

Findings

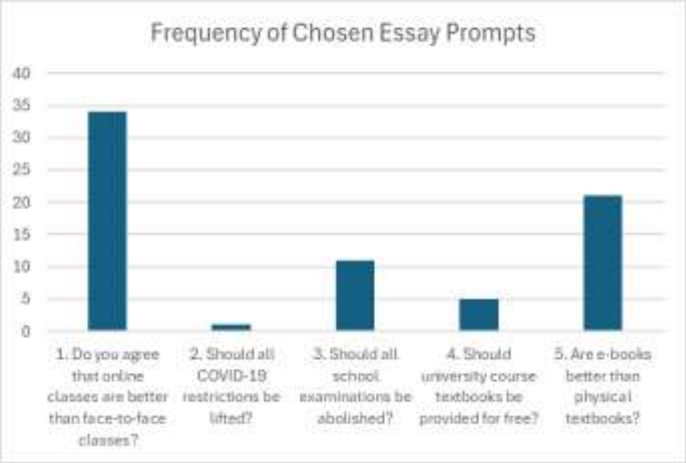

Figure 3: The Frequency of the Essay Prompts Chosen

The graph shows that the three most popular prompts are “Do you agree that online classes are better than face-to-face classes?”, “Are e-books better than physical textbooks?” and “Should all school examinations be abolished?”. The reason these three prompts are the most popular is that students relate to them the most due to personal experience.

Within the following table are the total averages for organisation, content and language, and vocabulary. Alongside the evaluator, grading was done with an AI application that was fed with the rubric so it could assess each essay accurately and within a short amount of time.

For the essay score comparison, the two-sample t-test was employed with the following null hypothesis tested:

H_0 - There is no significant difference between the Evaluator (human) scores and AI scores.

Table 4: T-Test Results

Criteria	Evaluator Mean	AI Mean	Differences	p-value
Organisation	7.58	4.93	2.65	0.00
Content	12.22	11.94	0.28	0.631
Language & Vocabulary	11.06	6.26	4.79	0.00
Total	**30.86**	**23.14**	**7.72**	**0.00**

The table displays the results of the t-test. Since the p-value of the “Organisation”, “Language & Vocabulary”, and “Total” essay scores is less than 0.05, the null hypothesis can be rejected, and therefore, there is a large difference between the scores assigned by the evaluator and the AI. However, the p-value for the “Content” scores is shown to be more than 0.05, and therefore fails to reject the null hypothesis, which means that there is no significant difference between the scores provided by the evaluator and the AI.

The scores assigned by the AI for both "Organisation" and "Language & Vocabulary" are significantly lower than the human evaluator's scores. As for "Content", both the AI and evaluator's scores are similar, which means that the AI's grading aligns closely with the evaluator in this area. When comparing the mean total scores of essays according to the evaluator and the AI, the AI consistently gives lower scores. We can conclude that the AI is stricter in grading "Organisation" and "Language & Vocabulary" as it focuses mainly on mechanics rather than overall fluency and coherence. While AI grading is more consistent overall, the evaluator demonstrates more variability in their scoring.

This rubric-based analysis evaluates students' essay performance across key areas: organisation, content, language use, and supporting evidence. Table 8 displays the average scores of the essays according to the evaluator and their corresponding ratings according to the rubric.

Table 5: Rubric-based Assessment of Essay Scores

Criteria	Average Score	Rating	Rubric's Definition
Organisation	7.58/12	Fair	"The writer demonstrates a modest ability to communicate ideas, with an average level of clarity in presenting thoughts. While key terms from the question are defined fairly, the arguments and logical reasoning supporting the premise remain modest, lacking strong development. Similarly, the conclusion presents a stance, though it is not firmly established or convincingly argued."
Content	12.22/20	Fair	"The writer provides relevant details and includes some additional, fair information to support their response. The content is organised fairly well, ensuring a logical flow of ideas. Additionally, the approach to the task demonstrates a fairly mature understanding, though there is room for deeper analysis and stronger argumentation."
Language & Vocabulary	11.06/18	Fair	The writer demonstrates some control over structure and grammar, though errors are present, with only about half of the sentences being grammatically correct. Vocabulary development is adequate, showing some variety in word choice. However, there are occasional issues with incorrect word forms, which affect clarity and precision in expression."

According to the evaluator, the essays scored fairly well across the board. They also stated that, "*Some introductions were missing clear thesis statements, and a few students who included thesis statements presented their main points in the body paragraphs in an order that did not* *align with the order outlined in their thesis. This, for me, impacted the coherence of their essays. Also, many students struggled to develop their main points fully. Elaboration was either lacking or, in some cases, repetitive and redundant. I think limited vocabulary can cause the students to not be able to elaborate on their ideas, and this hinders the overall clarity and comprehension of their writing. Most students are still struggling with the basic grammar concepts, as some of them are still making simple grammar mistakes like subject-verb agreement. Another noticeable mistake that I found was spelling. Lastly, the comprehension part. I had quite a hard time digesting what some students were trying to say in their essay."*

The AI-based assessment, on the other hand, provided a lot more details, including common issues in the writing and recommendations to improve the essays, which are seen in the table below.

Table 6: Common Issues Present In The Essays And Recommendations For Improvement

Criteria	Common Issues	Recommendations for Improvement
Organisation (Structure & Logical Flow)	● Weak transitions between ideas. ● Lack of a clear thesis statement. ● Disjointed paragraphs and abrupt conclusions.	● Teach topic sentences and logical progression. ● Use transitional phrases for smooth connections. ● Provide essay outlines to improve the structure.
Content (Depth & Relevance of Argument)	● Arguments lacked depth or strong supporting examples. ● Some essays were repetitive or went off-topic.	● Encourage critical thinking with justification for opinions. ● Teach counterarguments and rebuttals for stronger persuasion. ● Assign exercises to expand argument depth.
Language & Vocabulary (Grammar, Sentence Structure & Word Choice)	● Significant grammatical errors affecting clarity. ● Frequent verb tense and subject-verb agreement mistakes. ● Awkward phrasing due to direct translations. ● Limited vocabulary with repetitive word use.	● Grammar lessons focusing on verb tense, sentence structure, and subject-verb agreement. ● Reading assignments to expose students to academic writing. ● Peer editing to help identify and correct errors.

Additionally, the AI also listed additional errors that persist in the essays, as seen in Table 5.

Table 5 Additional Errors Observed from the Essays

Errors	Examples of Errors
Common Grammar & Sentence Structure Errors	● **Verb Tense Mistakes:** E.g., *"Students will difficult to understand"* → *"Students will have difficulty understanding"* ● **Subject-Verb Agreement Errors:** E.g., *"E-books is..."* → *"E-books are..."* ● **Preposition Misuse:** E.g., *"Face-to-face classes are better in online classes."* → *"Face-to-face classes are better than online classes."* ● **Run-On Sentences:** Long, unpunctuated sentences make comprehension difficult ● **Sentence Fragments:** Incomplete sentences, missing subjects or verbs
Vocabulary & Word Choice Issues	● **Awkward Phrasing:** Due to direct translation from students' native languages ● **Repetitive Words & Limited Vocabulary:** Lack of synonyms or varied sentence structures
Lack of Strong Supporting Examples	● General statements without concrete evidence. ● Improvements ● Require real-world examples in essays. ● Conduct debates and justification exercises.

According to Toulmin's Model, a strong persuasive argument consists of a claim, grounds, warrant, qualifier, backing and rebuttal. Using the model as a framework to analyse the essays, their contents are broken down to verify the presence of each element and how effectively they are presented. The figure below shows the frequency of Toulmin's Model argumentative elements in the essay samples.

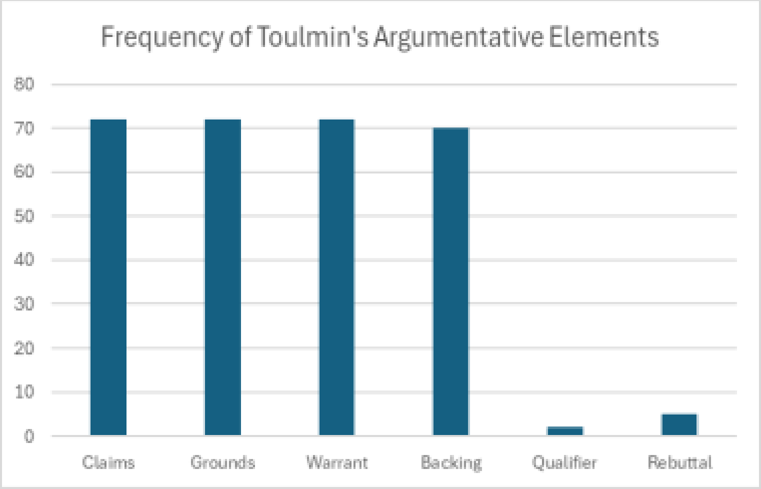

Figure 4: Frequency of Toulmin's Model Elements Throughout the 72 essays

The graph displays the frequency of each element in the essays. According to the table, Claims, Grounds and Warrants were found in all the essays (100%), and Backing is found in most of them (97.22%). However, there appears to be a lack of Qualifier and Rebuttal, which are the least used elements. The AI also summarises the common issues found associated with each element.

Each essay provided a clear and direct claim through a well-defined thesis statement. To support the claim, they provide logical grounds, even if many of them lack sufficient evidence. Despite these strengths, common weaknesses were also identified in the essays. The essays lacked qualifiers, as most of them made absolute statements. For example: "E-Books are always better". Only a handful of the essays addressed counterarguments or opposing perspectives. The essays lacked supporting evidence, as none included data, facts, expert opinions, or research studies. Instead, the arguments are only supported by the students' personal opinions, which are overused and repetitive.

Discussion

This study assessed the argumentative writing proficiency of university ESL students of various academic programmes using a rubric-based evaluation and Toulmin's Model. The findings highlight key weaknesses in organisation, argument depth, and language proficiency, reinforcing the need for targeted instructional support.

Based on the rubric-based evaluation, the essays demonstrate that students can perform reasonably well while developing their arguments. However, despite this, there are several glaring flaws in their composition.

First of all, they demonstrate weak organisational structure in their essays as they struggle with logical flow. There is a weak transition between their ideas as they try to elaborate on the details. A lack of a clear thesis statement is also observed in the essays. Finally, their paragraphs appear to be disjointed and end with abrupt conclusions. Toulmin's Model (Toulmin, 2003) analysis supports this, as many arguments lacked explicit warrants, making reasoning difficult to follow.

In terms of their content, their arguments lacked depth or strong supportive details. The students do not use examples, statistics, or facts to support their arguments. Instead, they only employ their own opinions. When attempting to make an argument, some essays present points repetitively, and there are even cases of essays presenting ideas that are off-topic from the original idea. Additionally, there is a severe lack of rebuttals in the essays, which shows that the majority of students can only recognise one side of the argument instead of both.

Finally, significant errors are observed in their grammar and vocabulary, which affect the clarity of the essays. There are frequent mistakes in verb tense and subject-verb agreement, as well as awkward phrasing resulting from direct translations, and a limited vocabulary, leading to repetitive use of words.

When comparing the grades assigned by the evaluator and AI, which both employ the same rubric, the AI appears to be far stricter when assigning scores for 'organisation' and 'language & vocabulary', as shown by the difference in mean scores. This demonstrates a discrepancy between grades assigned by human evaluators and those assigned by AI (Wetzler et al., 2024). However, the scores for 'content' are pretty similar to each other, showing a close agreement between the two scores. While the AI grading is stricter regarding grammar and vocabulary errors, both the AI and the evaluator agree that frequent language errors can impact comprehension, which in turn reduces the effectiveness of the argument.

ESL students struggle with argumentative structure, critical thinking, and language skills (Sabri et al., 2021; Joannes & AlSaqqaf, 2022). Similar studies (Osman & Januin, 2021) found that ESL learners often fail to include rebuttals and supporting evidence, reinforcing the need for explicit instruction. Discrepancies in AI vs. human grading (Almegren et al., 2024) highlight concerns about AI's effectiveness in evaluating coherence and argument depth. However, AI can still highlight and present a summary of the errors in the essays.

By incorporating Toulmin's Model into argumentative writing lessons, students can learn to structure their arguments by effectively incorporating backing and rebuttals. It is strongly recommended that students participate in workshops that focus on enhancing their language skills. Verb tense, sentence structure, and transitions, along with peer editing, can improve language skills. AI grading should not be disregarded entirely as it can be used to enhance, rather than replace, human evaluation, particularly for grammar and mechanics. The AI can still interpret the essay content, providing a summary of the errors and presenting recommendations for improvements. Finally, Students should be trained in finding external sources, statistics and expert opinions which can supplement their essays. This will require study and assigned reading on topics. For example, give students a topic to write about and assign them sources to find the evidence.

During the course of the study, a couple of limitations were noted. The first was the small sample size, as only 72 essays from a single university were analysed, thus limiting generalizability. Also, the AI applied a much stricter criterion for organisation and language, resulting in lower scores.

Conclusion

This study examined the persuasive writing proficiency of university ESL students using a rubric-based assessment and Toulmin's Model. Findings indicate that students struggle with organisation, depth of argumentation, and language proficiency, limiting the effectiveness of their persuasive essays. This can be seen in the Mean scores, which are average. When comparing the grading done by the AI and human evaluator, the AI scored the essays much lower, as seen in the mean scores.

While all of the essays are shown to use claim, grounds and warrant, with most employing backing, many essays lacked proper supporting evidence and rebuttals, relying heavily on personal opinions. Most do not use qualifiers, preferring to use absolute statements. Additionally, grammar and structural issues further affected clarity and coherence.

To improve ESL students' argumentative writing skills, several measures need to be enforced. First, explicit argumentative instruction that integrates Toulmin's Model into writing lessons should be implemented to help students practice constructing logical arguments with claims, backings, and rebuttals. Second, provide students with support in grammar and vocabulary. This can be achieved through workshops that focus on sentence structure, transitions, and common ESL errors, aiming to improve clarity and fluency. Once the students recognise their errors and learn how to fix them, they can improve their language skills. Next, AI-based grading tools should be utilised to provide prompt and straightforward feedback. Finally, evidence-based writing training encourages the students to incorporate external sources, real-world examples, and statistics into their essays.

This study offers invaluable insights into enhancing ESL argumentative writing instruction, emphasising the importance of structured writing frameworks, targeted pedagogical interventions, and the responsible use of AI in assessment. In the future, a larger study with an expanded sample size should be conducted, including students from multiple universities, to provide broader insights. An evaluation of Toulmin-based writing lessons should be done to assess the effectiveness of Toulmin's Model in the classroom. Finally, calibration can be performed to enhance the alignment between AI and human assessment.

A future study with a much larger and wider sample size could be conducted, focusing on more than just one university, as this study was primarily done within one university. It could also include universities in other states. This will be a challenging process, but help can be obtained from the lecturers of those institutions. This study can also be improved by recording each argumentative writing lesson and to gradually record the students' improvements.

Acknowledgement

We would like to express our sincere gratitude to all the students who participated in this inquiry and the evaluator who assisted in grading and analysing the essay samples. This study was also supported by a research grant (SLB2260) provided by the Research Management Centre of Universiti Malaysia Sabah.

References

Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., & Almegren, R. M. (2024). Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International*, 1-16.

Aziz, F. I. A., Salam, S. N., & Mat Said, S. B. (2023). Improving English Persuasive Writing in Malaysia: A Recommendation. *Asian Journal of Research in Education and Social Sciences*, *5*(2), 63–72.

Aziz, F.I.A. (2021). *The Persuasive Writing Features of Secondary School Students in Malaysia and Factors that Influence the Quality.* PhD Thesis. Universiti Teknologi Malaysia.

Aziz, F. I. A., & Ahmad, U. K. (2017). Persuasive Writing: How Students Argue. *Sains Humanika*, *9*(4-2), pp. 19–32.

Aziz, F. I. A., & Mat Said, S. B. (2019). Formulating a Modern Instructional Model to Help Improve the Persuasive Writing of Students. International Journal of Recent Technology and Engineering (IJRTE), *8*(3S2), pp. 924 – 927

Aziz, F. I. A., & Mat Said, S. B. (2020). The Success Factors That Impact the Quality of Students' Persuasive Essays. *Journal of Management & Muamalah*, *10*(1), pp. 30 – 45

Aziz, F. I. A., & Mat Said, S. B. (2020). Developing a persuasive writing model for secondary school. *Educational Research for Policy and Practice*, *19*(2), pp. 143–158.

Bahari, A. A., Kussin, H. J., Harun, R. N. S. R., Mohamed, M., & Jobar, N. A. (2021). The Limitations of Conducting Collaborative Argumentation When Teaching Argumentative Essays in Malaysian Secondary Schools. *Studies in English Language and Education, 8*(3), pp. 1111–1122.

Bouziane, K., & Bouziane, A. (2024). AI versus Human Effectiveness in Essay Evaluation. *Discover Education*, *3*(1), 201.

Campbell, Y. C., & Filimon, C. (2018). Supporting the argumentative writing of students in linguistically diverse classrooms: An action research study. Research in Middle Level Education, 41(1), 1-10.Available at: https://doi.org/10.1080/19404476.2017.1402408.

Dyrdal, G. M. (2021). *Scoring Rubrics – An Assessment Strategy to Promote Written English Competence in EFL-classrooms* (Master's thesis, Norwegian University of Science and Technology).

Flower, L., & Hayes, J. R. J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, *32*(4), pp. 365–387.

Ghulamuddin, N. J. A., Mohari, S. K. M., & Ariffin, K. (2021). Discovering writing difficulties of Malay ESL primary school level students. *International Journal of Linguistics and Translation Studies*, *2*(1), pp. 27-39.

Ghoneam, N (2023). The Effect of Applying the Toulmin Model on Enhancing Egyptian EFL Learners' Speaking Argumentative Skills, *12*(2), pp.119-145.

Goldman, J. (2019). Six High-Leverage Writing Practices for Teaching English Language Learners in English Language Arts. In de Oliveira, L. C., Obenchain, K. M., Kenney, R. H., & Oliveira, A. W. (Eds.). *Teaching the Content Areas to English Language Learners in Secondary Schools*. Springer, Cham, pp. 65–84.

Harris, K. R., & Graham, S. (2018). Self-regulated Strategy Development: Theoretical Bases, Critical Instructional Elements, and Future Research. InRedondo, R. F., Harris, K. & Braaksma, M. (Eds), *Design Principles for Teaching Effective Writing*, Leiden, Netherlands, Brill Publishers. pp. 119–151.

Hitchcock, D. (2017). Sound Reasoning on the Toulmin Model. In *On Reasoning and Argument*, pp. 371–387, Cham: Springer.

Indarti, D. (2025). Toulmin Argument Patterns in Asian EFL Learners Essays: Gender and Topic-Based Comparison. *Journal of Languages and Language Teaching*, *13*(3), pp. 1380-1392.

Joannes, R., & AlSaqqaf, A. (2022). The Effect of Analytic Text-based Writing Strategies on ESL Argumentative Writing among Malaysian Form-Six Students in Sabah, Malaysia: A Proposal. *In Proceedings International Conference on Teaching and Education (ICoTE) 3*(1) pp. 23–29.

Jonäll, K. (2024). Artificial Intelligence in Academic Grading: A Mixed-Methods Study.

Kunnel, J. G. (2021). *Raising Awareness About Task Assessment Rubrics in Task-based Language Teaching* (Doctoral dissertation, University of Calgary). PRISM Repository. http://hdl.handle.net/1880/113886

Manz, E. (2015). Representing Student Argumentation as Functionally Emergent from Scientific Activity. *Review of Educational Research*, *85*(4), pp. 553–590.

Marsaulina, R. M. (2024). Assessing Non-Native EFL Electrical Engineering Students' Argumentation Performance through Toulmin's Model-based Argumentative Genre Essay. *Jurnal Ilmiah Wahana Pendidikan*, *10*(3), 1001-1011.

Ministry of Education (MOE). (2013). Executive summary: Malaysia Education Blueprint 2013-2025 (Preschool to Post-Secondary Education). Putrajaya: Kementerian Pendidikan Malaysia.

Novianti, M. N. R. (March). Toulmin Model: A Strategy for Critical Thinking in Analytical Reading. In *International Conference on Educational Sciences and Teacher Profession (ICETeP, 2020),* pp. 296-302. Atlantis Press.

Osman, W. H., & Januin, J. (2021). Analysing ESL Persuasive Essay Writing using Toulmin's Model of Argument. *Psychol. Educ. J*, *58*, pp. 1810-1821.

Pei, Z., Zheng, C., Zhang, M., & Liu, F. (2017). Critical thinking and argumentative writing: Inspecting the association among EFL learners in China. English Language Teaching, 10(10), 31-42.Available at: http://doi.org/10.5539/elt.v10n10p31.

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. BMC Medical Education, 24(1), 962.

Ramli, N. H. L., Abdul Kadar, N. S., & Rafek, M. (2024). ESL Foundation Learners' Difficulties and Strategies Applied in Writing an Argumentative Essay Online. *ESTEEM Journal of Social Sciences and Humanities, 8*(1), pp. 1-13.

Rock, K. N. (2021). *Using Analytic Rubrics to Support Second Language Writing Development in Online Tasks* (Doctoral dissertation, University of Hawai'i at Manoa).

Sabri, S., Johan, S., Johan, A., Mohd, K., Fatinah, F., & Bahrn, S. (2021). Malaysian ESL Students' Perception of the Importance of Learning Argumentative Writing and Challenges Faced. *International Journal of Asian Social Science*. pp. 553-563

Sampson, V., Enderle, P., Grooms, J., & Witte, S. (2013). Writing to Learn by Learning to Write During the School Science Laboratory: Helping Middle and High School Students Develop Argumentative Writing Skills as They Learn Core Ideas. *Science Education*, *97*(5), pp. 643–670.

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025, March). Can AI Grade your Essays? A Comparative Analysis of Large Language Models and Teacher Ratings in Multidimensional Essay Scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 462-472).

Sundari, H., & Febriyanti, R. H. (2021). The analysis of Indonesian EFL argumentative writing using Toulmin's model: The structure and struggles of the learners. Scope: *Journal of English Language Teaching*, *5*(2), pp. 67-78.

Toulmin, S. (2003). *The Uses of Argument*. Cambridge University Press.

Toscu, S. (2023). Assessing Writing in EFL Context. *ELT Research Journal*, *12*(2), pp. 174–192.

UNESCO. (2019). Beijing Consensus on Artificial Intelligence and Education. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000368303 (Accessed 28 June 2025)

UNESCO. (2021). AI and Education: Guidance for Policy-Makers. https://unesdoc.unesco.org/ark:/48223/pf0000376709 (Accessed 28 June 2025)

UNESCO. (2023). Guidance for Generative AI in Education and Research. https://unesdoc.unesco.org/ark:/48223/pf0000386693 (Accessed 28 June 2025)

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education*, *12*(1), 20-32.

Vijaya Shetty, S., Guruvyas, K. R., Patil, P. P., & Acharya, J. J. (2022). Essay Scoring Systems using AI and Feature Extraction: A Review. *In Proceedings of the third international conference on communication, computing and electronics systems: ICCCES 2021,* pp. 45–57. Singapore: Springer Singapore.

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2024). Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. *Teaching of Psychology*