Stilistika: Jurnal Pendidikan Bahasa dan Seni
Volume 14, Nomor 2, Tahun 2026
P-ISSN: 2089-8460
E-ISSN: 2621-3338
Open Access: DOI: 10.59672/stilistika.v14i2.6280
https://ojs.mahadewa.ac.id/index.php/stilistika/index

A COMPARATIVE EVALUATION OF AI-GENERATED FEEDBACK: GEMINI VS CHATGPT IN ASSESSING EFL STUDENTS' GRAMMATICAL RANGE AND ACCURACY

I Made Agung Rai Antara[1*], Ni Putu Yunik Anggreni[2]
Department of Hotel Management, Triatma Mulya University[1], Department of Tourism, Triatma Mulya University[2]
Email: agung.rai@triatmamulya.ac.id[*], yunik.anggreni@triatmamulya.ac.id

A B S T R A K

Penelitian ini bertujuan untuk membandingkan kinerja kecerdasan buatan ChatGPT 5.3 dan Gemini pro 3.1 dalam menilai Grammatical Range and Accuracy (GRA) pada esai recount yang ditulis mahasiswa level Intermediate di Universitas Triatma Mulya. Penelitian ini menggunakan desain kualitatif komparatif dengan pendekatan analisis isi untuk mengeksplorasi, mengevaluasi, dan membandingkan kualitas umpan balik gramatikal yang dihasilkan oleh kedua sistem secara independen. Data penelitian berupa 15 esai mahasiswa tingkat menengah (Intermediate) di Universitas Triatma Mulya yang dianalisis berdasarkan jenis kesalahan, tingkat akurasi, dan rentang gramatikal. Hasil penelitian menunjukkan bahwa kedua model memiliki konsistensi tinggi dalam mengidentifikasi kesalahan utama, terutama pada aspek tense, struktur kalimat, dan kapitalisasi. Namun demikian, ditemukan perbedaan signifikan dalam kedalaman analisis dan tingkat sensitivitas terhadap kesalahan minor, di mana Gemini Pro 3.1 menunjukkan pendekatan yang lebih rinci dan berbasis aturan, sementara ChatGPT 5.3 cenderung memberikan penjelasan yang lebih sederhana dan mudah dipahami. Selain itu, Gemini menunjukkan kecenderungan evaluasi yang lebih ketat, sedangkan ChatGPT lebih moderat dalam klasifikasi tingkat akurasi dan rentang gramatikal. Temuan ini menunjukkan bahwa kedua sistem memiliki potensi yang kuat namun dengan orientasi yang berbeda, sehingga dapat digunakan secara komplementer dalam pembelajaran menulis bahasa Inggris.

Kata Kunci: *Kecerdasan Buatan, Evaluasi Gramatikal, ChatGPT, Gemini*

A B S T R A C T

This study aims to compare the performance of ChatGPT 5.3 and Gemini Pro 3.1 in assessing Grammatical Range and Accuracy (GRA) in recount essays written by intermediate-level students at Triatma Mulya University. The study employed a comparative qualitative design using a content analysis approach to explore, evaluate, and compare the quality of grammatical feedback independently generated by both artificial intelligence systems. The data consisted of 15 recount essays written by intermediate-level students at Triatma Mulya University, which were analyzed based on error types, accuracy levels, and grammatical range. The findings revealed that both models demonstrated a high level of consistency in identifying major grammatical errors, particularly in tense usage, sentence structure, and capitalization. However, significant differences were found in the depth of analysis and sensitivity to minor errors. Gemini Pro 3.1 tended to provide more detailed and rule-based feedback, whereas ChatGPT 5.3 offered explanations that were simpler and easier for students to understand. Furthermore, Gemini exhibited a stricter evaluative tendency, while ChatGPT adopted a more moderate approach in classifying grammatical accuracy and range. These findings suggest that both systems possess strong potential for grammatical assessment, albeit with different orientations, making them complementary tools in English writing instruction.

Keywords: Artificial Intelligence, Grammatical Assessment, ChatGPT, Gemini

Received : March, 2026	Revised : April, 2026	Accepted : May, 2026	Published : May, 2026

INTRODUCTION

The English as a Foreign Language (EFL) learning ecology has seen substantial changes due to advancements in artificial intelligence (AI) technology, especially regarding writing abilities. Current literature highlights that a primary focus of research is the integration of AI into automated writing evaluation (AWE) and feedback (Zhang & Liu, 2025). In educational practice, generative AI has evolved from being a passive tool into an active participant in the text-generation process. This shift requires pedagogical adjustments and raises ethical considerations in its application (Ruecker et al., 2014). Even while several studies indicate that AI-based feedback can enhance writing quality, challenges remain regarding its accuracy, dependability, and the risk of learners becoming overly dependent on technology (Zhai & Razali, 2023).

The learning environment also influences the necessity of using AI in EFL writing instruction. Due to their limited exposure to English outside of the classroom, EFL students typically write with poor linguistic complexity and accuracy (Bhowmik, 2021; Liao, 2018). However, structural issues like large class sizes and demanding workloads make it challenging for teachers to give detailed individual feedback (Anaktototy, 2023). Consequently, AI technology has increasingly been adopted as an alternative tool for providing writing feedback. However, because students often employ simple sentence structures to avoid error-free writing may actually hinder the development of grammatical range and accuracy (GRA) (Zhai & Razali, 2023).

Evaluating recount texts is particularly urgent because this genre requires precise chronological organization and consistent use of past tense structures. Research shows that EFL students frequently struggle with maintaining coherence between clauses, consistency in tense, and subject usage. These issues significantly impact overall quality of the text (Bhowmik, 2021; Dizon & Gayed, 2021; Fithriani, 2018; Liao, 2018). As a result, a feedback system that is both linguistically accurate and able to give learners understandable contextual explanations is needed. This study's theoretical foundation is the Automated Writing Evaluation (AWE) framework, which uses computational methods to mimic the automatic writing evaluation process (Cao et al., 2022; Lv et al., 2021). The key indicator utilized in this study is the Grammatical Range and Accuracy (GRA) parameter. Operationally, 'grammatical range' refers to the variety of sentence structures used, such as the mix of simple, compound, and complex sentences. Meanwhile, 'grammatical accuracy' refers to the frequency and severity of errors in aspects like tense usage, subject-verb agreement, and prepositions. These indicators are assessed based on standardized writing rubrics adapted for EFL contexts. The AI-generated response was qualitatively cross-checked against a human expert judgment, which functioned as the reference standard, to guarantee analytical validity.

A survey of the literature reveals that numerous studies have shown that AI models may achieve a high degree of reliability in evaluating objective grammar elements (Cao et al., 2022; Susanti, 2017). However, AI performance is still context-dependent, tends to deteriorate when no particular training has been given (zero-shot), and is less reliable when evaluating intricate and context-dependent elements (Dizon & Gayed, 2021; Xu & Li, 2023). Furthermore, the limitations of previous studies lie in their focus on applying single, older AI models in isolation. They have also heavily emphasized quantitative scoring rather than analyzing the qualitative pedagogical value of the feedback. Therefore, a significant research gap exists regarding how the newest generation of Large Language Models (LLMs) compare in providing pedagogically meaningful feedback for specific EFL genres like recount texts. This study addresses this gap

by comparing ChatGPT and Google Gemini. The rationale for selecting these two specific models is their current dominance as the most widely accessible LLMs for EFL students, despite utilizing different underlying computational architectures. This study is novel because it focuses not only on the accuracy of grammatical error detection but also on the pedagogical clarity of the feedback, specifically evaluating how explicit and comprehensible the AI's explanations are for learners.

In light of this, the purpose of this study is to evaluate and contrast the accuracy and consistency of ChatGPT and Google Gemini when evaluating the grammatical range and accuracy of recount essays written by EFL students. To achieve this, the study employed a qualitative-dominant approach by analyzing 15 recount texts written by EFL participants. The text data were collected and evaluated by both AI models simultaneously in April 2026 to minimize bias from algorithmic updates. The AI-generated feedback was then analyzed and compared against a human expert standard to measure both diagnostic accuracy and pedagogical clarity. This study is expected to contribute theoretically to the field of AWE by offering a fresh perspective on the comparative capabilities of modern LLMs in qualitative grammatical analysis. Practically speaking, the study's findings can serve as a guide for educators and educational establishments when choosing and using the most effective AI tools to enhance writing evaluation and instruction in the digital era.

METHOD

This study uses a qualitative content analysis approach with a comparative qualitative design. The goal is to examine, evaluate, and compare the quality of Grammatical Range and Accuracy (GRA) input generated by Google Gemini and ChatGPT. The comparison centers on the feedback's instructional value for students learning English as a foreign language (EFL). Instead of concentrating on differences in statistics scores, it specifically looks at the accuracy of error detection and the depth of grammatical explication. Purposive sampling was used to pick 15 recount essays on the theme of "Unforgettable Experience" for the research data. The concept of qualitative data saturation provides academic justification for the sample size of 15 texts, despite the fact that it may seem modest. Without the need for a bigger dataset, this figure offers enough depth for a micro-analytical investigation of AI feedback patterns. Fifteen students from Triatma Mulya University wrote these essays in order to explain the research context. The participants, who were between the ages of 18 and 19, had an intermediate level of English competence, which made them a perfect group to examine typical GRA mistakes in recount writing.

The students' essay writings and the comments generated by both AI systems were the two primary sources of data collected utilizing document analysis techniques. All data collecting was carried out methodically in April 2026 to reduce any AI biases resulting from model changes or variances in generating outputs. Each essay was then thoroughly examined by both AI systems. The instructions given to both systems were standardised using the following prompt: “Act as an English writing tutor. Evaluate the following EFL student's recount essay titled “Unforgettable Experience”. Focus ONLY on Grammatical Range and Accuracy (GRA), including verb tense, sentence structure, agreement, articles, and prepositions. Identify each grammatical error.Classify the type of error (e.g., tense, agreement, article, etc.). Explain why it is incorrect. Provide the corrected version. Present your analysis in a table format with columns: Error, Type, Explanation, Correction. Do NOT comment on content, vocabulary, or organization.” (A single-trial prompt design was implemented to ensure output consistency across both AI platforms). Data analysis was carried out by adapting an interactive model from (Miles et al., 2014), which consists of three primary steps: condensing data, presenting facts, and drawing conclusions. In order to focus the analysis on GRA features, the feedback output from both AIs was streamlined during the data condensation stage by

removing unnecessary components. The data that is, the analytical outcomes from every AI system was then shown as a comparative matrix. Qualitative analysis is then conducted based on two main indicators, namely: (1) grammatical correction accuracy, measured by the alignment between the AI's corrections and human expert benchmark provided by a practitioner with 15 years of EFL teaching experience. and (2) pedagogical clarity, reflecting the extent to which the explanations provided by the AI can be understood and utilised by students. This rigorous analysis process yields an in-depth interpretation of the strengths and limitations of each system in providing GRA-based feedback, which is subsequently synthesised into a comprehensive narrative conclusion.

RESULTS AND DISCUSSION

Results

The study's findings compare the grammatical remarks produced by ChatGPT and Google Gemini to assess the Grammatical Range and Accuracy (GRA) of recount essays turned in by EFL students. Before delving deeply into the comparative results, it is necessary to establish the methodological basis for the percentages displayed in the following tables. Instead of a simple diagnostic accuracy score, the percentages reflect the frequency of certain error occurrences detected over the full unit of analysis (n = 15 essays). The calculation formula was (number of essays with a specific type of error/total number of essays, n=15) × 100. For instance, a 100% frequency in "verb tense" indicates that this specific issue was correctly discovered in all 15 recount documents. This explicit computation ensures that the data classification is accurate and reproducible. It should be mentioned that the human expert's assessment serves as the ultimate "ground truth" to confirm if the AI's detection and adjustments are real or imagined, rather than being displayed as a separate variable in the compared tables.

To provide particular evidence of how both models operate and to highlight their differences in pedagogical clarity, Table 1 presents a real example of the feedback generated by both AI algorithms, derived from Student Data 1.

Table 1. Authentic Example of AI Feedback Generation (Extracted from Student Data 1)

No	Aspect	Content
1	Student's Original Sentence	"because we are missed them so much, i goes there..."
2	Human Expert's Correction	Correction: "because we missed them so much, I went there..." Note: The student mixed present/passive ("are missed") and present ("goes") in a past recount context. Both verbs need to be converted to active simple past.
3	ChatGPT's Correction & Explanation	**Error 1:** "we are missed" (Type: Tense). **Explanation:** Mixing present "are" with past "missed". **Correction:** "because we missed". **Error 2:** "i goes" (Type: Verb tense/agreement). **Explanation:** Past tense required. **Correction:** "I went".
4	Gemini's Correction & Explanation	**Error 1:** "we are missed" (Type: Verb Tense/ Voice). **Explanation:** "Are missed" is present tense and passive voice. Because the student is recounting a past event, it requires the simple past tense in the active voice. **Correction:** "we missed". **Error 2:** "goes" (Type: Verb Tense). **Explanation:** The essay is a recount of a past event, so the past tense form of the verb "go" must be used. **Correction:** "went".
	Comparative Analysis	This genuine excerpt demonstrates how the teaching delivery differs. ChatGPT uses straightforward teaching language. (e.g.,

5		"Mixing present with past"). Gemini, on the other hand, shows a more rigorous, formal grammar-checking architecture by using extremely technical, metalinguistic language (such as "passive voice," "active voice," and "recounting a past event").

Table 1 shows significant distinctions between the two AI models' instructional delivery strategies, going beyond a cursory comparison. From a conceptual standpoint, ChatGPT serves as a scaffolding tutor; it recognises mistakes and offers fixes in simple, understandable conversational language (e.g., "Mixing present and past tenses"). In keeping with the tenets of meaning-focused learning, this method lessens the cognitive load on students. Gemini, on the other hand, is a rigid, rule-based evaluator. Although very accurate in terms of diagnosis, this model uses sophisticated metalinguistic terminology (such as "passive voice" and "active voice"), which necessitates that students have a greater degree of prior grammatical understanding.

Table 2.Analysis of Error Detection and Occurrence Pattern

No	Error Category	Occurrence Pattern	ChatGPT Detection	Gemini Detection
1	Verb tense and verb form	Very Frequent	100%	100%
2	Sentence structure (Run-ons/Fragments	Frequent to Very frequent	100%	100%
3	Capitalization	Very frequent	100%	100%
4	Spelling	Frequent	93%	93 %
5	Article and noun agreement	Moderate to high	80%	87%
6	Preposition errors	Moderate to high	73%	80%
7	Lexical/Word form	Moderate	80%	87%
8	Mechanics (Comma splice/punctuation	Very frequent	Limited	100 %

Table 2 indicates that while both models exhibit flawless detection rates (100%) for macro-level errors such verb tenses and basic sentence structure, their analytical depth exhibits notable micro-level variance. When it comes to more specific linguistic mechanics, including the utilisation of articles (87%), prepositions (80%), and improper comma usage or comma splices (100%), Gemini consistently exhibits a higher level of sensitivity. Analytically, this discrepancy highlights how the two models' underlying technology architectures differ. Gemini's data extraction technique makes it extremely aware of punctuation and structural limitations, and it tends to parse English similarly to mathematical code. On the other hand, ChatGPT's Generative Pre-trained Transformer architecture accounts for its comparatively lower detection rate of mechanical defects, including comma splices. Rigid syntactic correctness is subordinated to conversational flow and overall comprehensibility in this architecture. This finding implies that Gemini outperforms ChatGPT in thorough and in-depth proofreading tasks for English as a foreign language (EFL) learners, while ChatGPT tends to be more forgiving of little mistakes that do not impair meaning.

Table 3. Classification of Grammatical Range and Accuracy

No	Parameter	Classification Level	ChatGPT	Gemini	Description of Discrepancy
	Grammatical Accuracy	Low	40%	47%	Gemini places more essays in the lowest accuracy level

1		Moderate-Low	33%	33%	Both models show equal distribution
		Moderate	20%	13%	Chat GPT assigns more essays on the moderate level
		Moderate high	7%	7%	Both models identify few high-accuracy texts
2	Grammatical Range	Very limited	13%	13%	Both models identify severe lack of variance
		Limited	33%	40%	Gemini is stricker in evaluating structural variety
		Moderate	47%	40%	ChatGPT is more lenient, placing more in moderate
		Moderate-Good	7%	7%	Both shows equal performance at the highest level

As detailed in Table 3, the classification distribution shows that the two systems' evaluation biases differ significantly. Gemini's evaluation findings are constantly skewed toward lower categories (e.g., 40% in Limited Range and 47% in Low Accuracy). This suggests the existence of a "penalty-based" evaluation system, in which Gemini adds up each small mechanical error (as seen in Table 2), ultimately reducing the final score. On the other hand, a greater proportion of essays are classified as "Moderate" by ChatGPT (20% for Accuracy and 47% for Grammatical Range). From an analytical standpoint, this implies that ChatGPT uses a comprehensive evaluation rubric with a "fit-to-intended-meaning" focus. ChatGPT is designed to recognise when a student's recount essay effectively communicates its chronological story in spite of grammatical errors. This shows an important conclusion: AI systems are not totally impartial evaluators; rather, their assessment behaviour reflects particular pedagogical philosophies, whether communicative (as in ChatGPT) or form-focused (as in Gemini).

Table 4. Synthesis of Analytical Focus and Overal AI Performance

No	Dimension	ChatGPT Characteristics	Gemini Characteristics
1	Error Coverage and Precision	High (focuses on major communicative errors)	Very high (Exhaustive and rule-oriented)
2	Explanation style	Simple, accessible and conversational	Technical, formal, specific metalinguistic terms
3	Structural analysis	Moderate (prioritizes flow)	Detailed (Deep syntactic breakdown)
4	Strictness Level	Moderate (forgiving of minor mechanical flaws)	High (highly critical of punctuation /form)
5	Overall Tendency	Interpretive and Learner-Friendly	Analytical and Evaluative

The extensive behavioural profiles of both Large Language Models (LLMs) are included in Table 4. The training approaches that support each system are closely related to a thorough interpretation of these characteristics. The Reinforcement Learning from Human Feedback (RLHF) mechanism, which teaches the AI to function as a helpful and sympathetic helper rather than just a strict machine, is directly responsible for ChatGPT's "interpretive" and "learner-friendly" inclinations. On the other hand, Gemini's "Analytical" profile with a "Very High" strictness level corresponds with its algorithmic instructions, which are made especially for accurate fact-checking and information extraction.This suggests that using ChatGPT mimics the experience of interacting with a helpful human tutor who continuously encourages students to produce language in the context of English as a Foreign Language (EFL) writing instruction.

Gemini, on the other hand, mimics the process of sending a draft to an extremely rigorous computerised grammar checker. As a result, depending only on one system will give students an inadequate educational experience.

Discussion

This study examines how well ChatGPT and Gemini perform when evaluating Grammatical Range and Accuracy (GRA) in recount essays written by EFL students. It also finds parallels and discrepancies between the two systems' analytical behaviors. When it comes to identifying the primary categories of grammatical errors, especially those involving tense usage, verb forms, sentence structure, and capitalization, both models show a very high level of consistency. These results show that both algorithms are highly reliable in identifying simple grammatical faults that are governed by rules. Additionally, the frequency of verb-related mistakes in all essays indicates that EFL learners continue to face significant difficulties with consistent tense usage.This is especially noticeable in recount writings, where the past tense must be used steadily and consistently. As a result, both approaches might be regarded as successful in pinpointing the most basic grammar problems in students' writing. These results suggest that both systems are very reliable in identifying simple, rule-governed grammatical problems, which is consistent with earlier studies demonstrating the speed and accuracy with which automated writing assessment systems can identify faults in grammar, spelling, and punctuation (Chen & Pan, 2022; Mariappan et al., 2022). Additionally, the discovery that grammar is the primary challenge EFL students have in academic writing supports the frequency of verb-related errors (Asnas & Hidayanti, 2024; Tambunan et al., 2022).

The two models show notable disparities in terms of the depth of analysis and sensitivity to errors, despite their similarity in the extent of error identification. A greater variety of grammatical mistakes, such as the usage of articles, prepositions, word forms, and mechanical elements like punctuation and comma splices, are reliably recognized by Gemini. Gemini takes a more rule-based approach that emphasizes grammatical form correctness, as evidenced by the usage of specific grammatical terms such "auxiliary verb," "clause error," and "gerund." This feature is consistent with the AI's superior diagnostic function in accurately identifying and classifying mistakes at the linguistic mechanics surface level (Mariappan et al., 2022; Sanosi, 2022; Tambunan et al., 2022). On the other hand, ChatGPT tends to offer more comprehensive and easily comprehensible explanations and exhibits a little less sensitivity to small mistakes. This method emphasizes "fit-to-intended-meaning" and student comprehension, reflecting a more interpretative perspective (Faisal & Carabella, 2023). ChatGPT seems to concentrate more on the prognosis and communicative aspects rather than strict syntactic constraints, offering feedback that is more comprehensive and learner-friendly a trend commonly seen in Large Language Models when assessing written material (Danping, 2024; Mariappan et al., 2022). This distinction demonstrates how the two systems' analytical traits and degrees of specificity differ when assessing a text's grammatical quality.

The accuracy ratings and grammatical scope that the two systems assign also show differences. More essays fall into the low accuracy and limited range categories because to Gemini's tendency to apply a harsher judgment. This suggests that Gemini prioritizes grammatical accuracy more strictly in its evaluation process, which is consistent with the traits of automated systems that place a strong emphasis on surface constraints and "grammatical form correctness (Faisal & Carabella, 2023; Sanosi, 2022). On the other hand, certain writings are given a higher score than in Gemini's evaluation because ChatGPT typically offers more mild categories. This tendency implies that ChatGPT takes a more forgiving evaluation stance toward mistakes, as long as the meaning is still understandable. The conclusion that assessments based on large language models (LLMs) frequently give priority to "fit-to-intended-meaning" and the author's communicative aim is supported by this grading pattern (Faisal & Carabella,

2023; Mun, 2024). Additionally, ChatGPT-style systems frequently prioritize increasing the text's overall fluency, which occasionally results in a lack of rigor in syntactic examination (Danping, 2024). The form-based approach, which emphasizes accuracy, and the communicative approach, which emphasizes comprehensibility, are the two primary methods of language assessment that account for this discrepancy.

Theoretically, these results align with important ideas in writing evaluation and second language acquisition (SLA). The classic contrast between communicative language instruction (CLT) and accuracy-oriented techniques is seen in the contrasts between the two models. Gemini seems to be more in line with the Form-Focused Instruction (FFI) approach, where learning is centered on explicit attention to grammatical rules, structural accuracy, and error elimination (Alsariera & Alsaraireh, 2024; Saleem et al., 2025). Additionally, the Noticing Hypothesis is highly supported by Gemini's explicit and thorough feedback; the system uses a cycle of repeated corrections to assist learners become aware of particular linguistic qualities (Jabsheh, 2024). On the other hand, ChatGPT exhibits a communicative strategy that prioritizes fluidity, meaning building, and communicating the writer's goal (Chick, 2025; Mills et al., 2025). For students with lower skill levels in particular, a more straightforward description of ChatGPT that emphasizes higher-level conversation can lessen cognitive burden and make the learning process more approachable. Additionally, this method teaches students to prioritize the presentation of meaning and helps them resist the pressures of strict linguistic standardization (Crompton et al., 2024; Zhou, 2023). Therefore, each method has advantages of its own, indicating that the most complete pedagogical approach to using AI in language learning is a practical integration of form-based and meaning-based assessment (Aljuaid, 2024; Saleem et al., 2025).

These findings have important pedagogical and practical ramifications for the ecology of EFL learning. The findings imply that teachers shouldn't use AI tools consistently since their technological prejudices have a direct impact on student learning. Gemini is more appropriate for learning situations that call for extensive grammatical analysis and a high level of precision, like advanced evaluation or the revision stage. However, in the early phases of learning, when students need concise, easy-to-understand explanations, ChatGPT works better. A balance between grammatical correctness and comprehensibility can be achieved by combining the use of both systems. This method is in line with contemporary language acquisition theories that combine training that is both form-focused and meaning-focused. The learning process can become more efficient and flexible to meet the demands of students by utilizing the advantages of each system. Overall, both systems show great promise for evaluating Grammatical Range and Accuracy, but they reflect distinct pedagogical philosophies. While ChatGPT serves as a more learner-friendly, communicative-focused feedback giver, Gemini is a precise, rule-based assessor. As a result, the two systems should be seen as complementing resources rather than as tools that are mutually exclusive. The greatest benefits can be obtained by incorporating both into EFL writing training, since this will enhance both the development of communication skills and grammatical precision. These results have significant ramifications for the future development of technology-based teaching methods. In terms of student performance, both models agreed that the participants' grammatical accuracy and range are mainly restricted to moderate, often without consistent structural control.

CLOSING

Conclusion

This study shows that ChatGPT and Google Gemini are both very sufficient at spotting significant grammatical mistakes in the recount essays of EFL students, especially when it comes to capitalisation, sentence form, and tense usage. Both models consistently identified these overt rule breaches across the dataset, demonstrating that verb-related inconsistencies

continue to be the key obstacle for EFL learners. As a result, both AI technologies are quite useful for basic Grammatical Range and Accuracy (GRA) assessments. The study finds notable differences in their analytical depth and feedback sensitivity despite this shared capability. Gemini has a rigid, rule-based structure and is particularly sensitive to little mistakes in punctuation, articles, and prepositions. Additionally, it uses sophisticated metalinguistic terms, which leads to a strict assessment that places more essays in lower competency bands. On the other hand, ChatGPT takes a meaning-focused, interpretive stance. In general, it assigns more moderate accuracy scores and prioritises overall comprehensibility by providing simplified, learner-friendly explanations. In terms of student performance, both models agreed that the participants' grammatical accuracy and range are mostly restricted to moderate, often without consistent structural control.In the end, the results show that although both models have a great deal of potential for teaching EFL writing, they represent different pedagogical philosophies. Gemini works best when used as an analytical, exact instrument for sophisticated grammatical analysis. On the other hand, ChatGPT is a user-friendly tutor that is perfect for helping students understand. Thus, this study finds that the most thorough and efficient method for automated writing evaluation is to deliberately integrate both models, striking a balance between Gemini's diagnostic accuracy and ChatGPT's pedagogical clarity.

Suggestion

Teachers are recommended to incorporate ChatGPT and Google Gemini as supplementary aids in EFL writing instruction based on the findings. Teachers could carefully integrate both models to optimise pedagogical benefits: using Gemini for thorough, in-depth syntactic analysis and ChatGPT for easily comprehensible, learner-friendly explanations of foundational faults. Additionally, for self-directed practice, students are urged to use this dual-system method, first using ChatGPT to understand simple corrections then contacting Gemini for more complex grammatical precision. To avoid becoming overly dependent on automated feedback, students must continue to use their critical thinking abilities. It is advised that future studies increase the sample size and investigate a variety of genres, including descriptive and argumentative texts. Future research should look into how AI affects writing skills over the long run and expand the evaluation criteria beyond grammatical correctness to include coherence, cohesiveness, and lexical richness in EFL contexts.

REFERENCE

Aljuaid, H. (2024). *The Impact of Artificial Intelligence Tools on Academic Writing Instruction in Higher Education: A Systematic Review*. https://doi.org/10.31235/osf.io/ph24v

Alsariera, A. H., & Alsaraireh, M. Y. (2024). Advancing EFL Writing Proficiency in Jordan: Addressing Challenges and Embedding Progressive Strategies. *International Journal of Arabic-English Studies*. https://doi.org/10.33806/ijaes.v24i2.664

Anaktototy, K. (2023). Interplaying Reading and Writing in ESL/EFL: A Literature Review of Strategies for Indonesian Teachers. *Elsya Journal of English Language Studies*, *5*(1), 107–121. https://doi.org/10.31849/elsya.v5i1.9994

Asnas, S. A. M., & Hidayanti, I. (2024). Uncovering EFL Students' Frequent Difficulties in Academic Writing and the Coping Strategies: The Case of a College in Indonesia. *Journal on English as a Foreign Language*, *14*(1), 124–151. https://doi.org/10.23971/jefl.v14i1.7472

Bhowmik, S. (2021). Writing Instruction in an EFL Context: Learning to Write or Writing to Learn Language? *Belta Journal*, *5*(1), 30–42. https://doi.org/10.36832/beltaj.2021.0501.03

Cao, S., Zhou, S., Luo, Y., Wang, T., Zhou, T., & Xu, Y. (2022). A Review of the ESL/EFL

Learners' Gains From Online Peer Feedback on English Writing. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1035803
Chen, H., & Pan, J. (2022). Computer or Human: A Comparative Study of Automated Evaluation Scoring and Instructors' Feedback on Chinese College Students' English Writing. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1). https://doi.org/10.1186/s40862-022-00171-4
Chick, J. C. (2025). *Writing With AI at the Margins: Student Voice and Authenticity at a Minority-Serving Institution*. https://doi.org/10.21203/rs.3.rs-8427622/v1
Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English Language Teaching: Affordances and Challenges. *British Journal of Educational Technology*, *55*(6), 2503–2529. https://doi.org/10.1111/bjet.13460
Danping, D. (2024). *Tapping Into the Pedagogical Potential of infinigoChatIC: Evidence From iWrite Scoring and Comments and Lu &Amp;amp; Ai's Linguistic Complexity Analyzer*. https://doi.org/10.31235/osf.io/xnrtz
Dizon, G., & Gayed, J. M. (2021). Examining the Impact of Grammarly on the Quality of Mobile L2 Writing. *The Jalt Call Journal*, *17*(2), 74–92. https://doi.org/10.29140/jaltcall.v17n2.336
Faisal, F., & Carabella, P. A. (2023). Utilizing Grammarly in an Academic Writing Process: Higher-Education Students' Perceived Views. *Journal of English Language Teaching and Linguistics*, *8*(1), 23. https://doi.org/10.21462/jeltl.v8i1.1006
Fithriani, R. (2018). Cultural Influences on Students' Perceptions of Written Feedback in L2 Writing. *Journal of Foreign Languange Teaching and Learning*, *3*(1). https://doi.org/10.18196/ftl.3124
Jabsheh, A.-A.-H. M. M. (2024). Relevancy and Outlook of the Technology-Enhanced Education Within Digital Contents, Resources and Tools. *Ijmer*, *3*(1), 24–34. https://doi.org/10.32996/ijmer.2024.3.1.4
Liao, F.-Y. (2018). Prospective ESL/EFL Teachers' Perceptions Towards Writing Poetry in a Second Language: Difficulty, Value, Emotion, and Attitude. *Eurasian Journal of Applied Linguistics*, *4*(1), 1–16. https://doi.org/10.32601/ejal.460583
Lv, X., Ren, W., & Xie, Y. (2021). The Effects of Online Feedback on ESL/EFL Writing: A Meta-Analysis. *The Asia-Pacific Education Researcher*, *30*(6), 643–653. https://doi.org/10.1007/s40299-021-00594-6
Mariappan, R., Tan, K. H., Yang, J., Jian, C., & Chang, P. K. (2022). Synthesizing the Attributes of Computer-Based Error Analysis for ESL and EFL Learning: A Scoping Review. *Sustainability*, *14*(23), 15649. https://doi.org/10.3390/su142315649
Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). SAGE Publications.
Mills, N., Hok, H., Dressen, A., & Veillas, Q. (2025). The Design and Evaluation of an Interactive AI Companion for Foreign Language Writing. *Foreign Language Annals*, *58*(1), 40–69. https://doi.org/10.1111/flan.12790
Mun, C. (2024). EFL Learners' English Writing Feedback and Their Perception of Using ChatGPT. *Stem Journal*, *25*(2), 26–39. https://doi.org/10.16875/stem.2024.25.2.26
Ruecker, T., Shapiro, S., Johnson, E. N., & Tardy, C. M. (2014). Exploring the Linguistic and Institutional Contexts of Writing Instruction in TESOL. *Tesol Quarterly*, *48*(2), 401–412. https://doi.org/10.1002/tesq.165
Saleem, T., Saleem, A., & Aslam, D. M. (2025). Integrating AI in Pakistani ESL Classrooms: Teachers' Practices, Perspectives, and Impact on Student Performance. *Plos One*, *20*(9), e0333352. https://doi.org/10.1371/journal.pone.0333352
Sanosi, A. (2022). To Err Is Human: Comparing Human and Automated Corrective Feedback. *Information Technologies and Learning Tools*, *90*(4), 149–161.

https://doi.org/10.33407/itlt.v90i4.4980
Susanti, A. (2017). Teachers' Corrective Feedback on Students' L2 Writing: State of the Art. *Abjadia International Journal of Education*, *2*(2), 81–94. https://doi.org/10.18860/abj.v2i2.5364
Tambunan, A. R. S., Andayani, W., Sari, W. S., & Lubis, F. (2022). Investigating EFL Students' Linguistic Problems Using Grammarly as Automated Writing Evaluation Feedback. *Indonesian Journal of Applied Linguistics*, *12*(1), 16–27. https://doi.org/10.17509/ijal.v12i1.46428
Xu, Q., & Li, P. (2023). Computational Modeling of Language Learning in the Era of Generative Artificial Intelligence: A Response to Open Peer Commentaries. *Language Learning*, *73*(S2), 83–94. https://doi.org/10.1111/lang.12605
Zhai, X., & Razali, A. B. (2023). Triple Method Approach to Development of a Genre-Based Approach to Teaching ESL/EFL Writing: A Systematic Literature Review by Bibliometric, Content, and Scientometric Analyses. *Sage Open*, *13*(1). https://doi.org/10.1177/21582440221147255
Zhang, S., & Liu, X. (2025). Learner Emotions in AI-assisted English as a Second/Foreign Language Learning: A Systematic Review of Empirical Studies. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1652806
Zhou, Y. (2023). The Effectiveness of Automated Written Corrective Feedback on L2 Learners' Revision Outcomes: A Case for ChatGPT. *International Journal of New Developments in Education*, *5*(25). https://doi.org/10.25236/ijnde.2023.052511