Volume 1 Issue 2 Year 2025 Pages 21-41
e–ISSN 3090–9406 | DOI: 10.70152
https://journal.akademimerdeka.com/ojs/index.php/duites

ChatGPT's Handling of L2 Learners' Fossilized Errors: A Linguistic Evaluation

Zahratun Nufus[1*], Saleman Mashood Warrah[2]

[1] STAI Rasyidiyah Khalidiyah (Rakha) Amuntai, Indonesia
[2] Kwara State University, Malete, Nigeria
[*]Corresponding author's email: zahratun918@gmail.com
DOI: https://doi.org/10.70152/duties.v1i2.219

Abstract: This study investigates ChatGPT's capacity to address fossilized grammatical errors in English as a Foreign Language (EFL) learners' academic writing. Through a mixed-methods design, a controlled corpus of 500 hypothetical sentences containing persistent error types, such as verb tenses, articles, prepositions, and non-idiomatic expressions, was submitted to ChatGPT-4. Quantitative analysis evaluated correction accuracy using standard metrics (precision, recall, F-score), while qualitative content analysis assessed the pedagogical appropriateness and consistency of ChatGPT's feedback. Results showed high accuracy in correcting rule-based structures (e.g., subject-verb agreement), but significantly lower performance for context-sensitive and fossilized errors. While ChatGPT often provided clear corrections, its feedback frequently lacked explanatory depth, contextual sensitivity, and scaffolding necessary for promoting learner noticing and long-term acquisition. These findings suggest that although ChatGPT can effectively support surface-level proofreading, it cannot fully substitute the role of human instructors in addressing deeply ingrained L2 errors. The study emphasizes the importance of explainable AI, AI literacy, and hybrid instructional models that combine technological efficiency with pedagogical intentionality. It offers implications for educators, curriculum developers, and AI tool designers seeking to integrate language models into second language acquisition contexts.

Keywords: academic writing, ChatGPT, error correction, EFL learners, fossilization

INTRODUCTION

The global expansion of English as a Foreign Language (EFL) education has elevated academic writing to a central skill for learners aiming to participate in international academic and professional contexts. Mastery of academic writing in a second language (L2) extends beyond vocabulary knowledge and basic syntactic structures, requiring learners to internalize complex grammar and sophisticated rhetorical conventions (Canagarajah, 2024; Kormos, 2023). However, despite years of instruction and practice, many L2 learners continue to produce persistent, recurring errors—referred to as *fossilized errors* (Albelihi & Al-Ahdal, 2024). Fossilization occurs when incorrect

language forms become ingrained habits that are resistant to correction, even at advanced proficiency levels (Long, 2015). These errors often span multiple linguistic domains, including phonology, morphology, syntax, and lexis, and pose particular challenges in academic writing, where precision and accuracy are paramount. The causes of fossilization are multifaceted. Common factors include negative transfer from the first language (L1), inconsistent or insufficient instruction, limited opportunities for meaningful corrective feedback, and cognitive entrenchment of erroneous forms (Albelihi & Al-Ahdal, 2024). For example, differences in grammatical structures between L1 and L2—such as pronoun distinctions in English versus Spanish or vowel length in English versus Chinese—can lead to persistent, uncorrected patterns of error. Given the persistent nature of fossilized errors and the limitations of traditional feedback mechanisms, educators and researchers have increasingly turned to technology-driven solutions to enhance the efficacy and immediacy of corrective feedback.

Recent advances in Artificial Intelligence (AI), particularly the development of Large Language Models (LLMs) like ChatGPT, have introduced new possibilities in language education. Trained on extensive corpora, ChatGPT can generate fluent, contextually relevant responses and provide immediate feedback, offering practical benefits in writing instruction (Xiao & Zhi, 2023). Its utility has been explored in areas such as genre classification, writing assistance, and error correction, with findings suggesting potential advantages over traditional Automated Writing Evaluation (AWE) systems (Bucol & Sangkawong, 2025). Unlike earlier AWE tools, ChatGPT is capable of producing tailored, conversational feedback that aligns more closely with human interaction.

Despite these advancements, a significant gap remains in understanding how effectively ChatGPT addresses *fossilized* grammatical errors—a subset of learner errors known for their resistance to standard correction methods. While ChatGPT has demonstrated proficiency in general grammatical error correction, fossilized errors pose a unique challenge, often requiring more than surface-level identification and correction. Effective second language feedback must not only provide correct forms but also foster learner awareness, facilitate self-repair, and offer developmentally appropriate explanations that support lasting linguistic change (Zhang & Zhang, 2023). However, current AI models, including ChatGPT, often function as "black boxes," offering corrections without transparent rationales, which may undermine their pedagogical effectiveness (Kos & Mažgon, 2025). This study aims to critically evaluate ChatGPT's capacity to (1) accurately identify and correct fossilized grammatical errors in EFL academic writing, and (2) deliver pedagogically appropriate and consistent feedback across different grammatical structures. By focusing on fossilized errors common in EFL contexts—such as verb forms, tense usage, article choice, prepositions, subject-verb agreement, and voice—the study isolates areas of persistent difficulty for learners. To ensure consistency and control, the analysis employs a hypothetical corpus of academic texts embedded with fossilized errors, benchmarked against expert-annotated corrections.

The implications of this research are twofold. From a pedagogical perspective, it offers educators insights into how ChatGPT might be integrated into instruction, particularly in

hybrid models that pair AI tools with human guidance. For learners, the findings may help in developing critical digital literacy skills and more strategic use of AI feedback. From an AI development perspective, the study highlights performance boundaries and suggests pathways toward more *explainable AI* in grammar correction—models that not only correct but also elucidate. In doing so, this research contributes to ongoing discussions about AI's evolving role in language learning—not as a replacement for teachers, but as a complementary tool that, when properly understood and applied, can enhance both instruction and learner autonomy.

LITERATURE REVIEW

Theoretical and Conceptual Framework

The phenomenon of fossilization—where persistent language errors become resistant to correction despite continued exposure and instruction—has been a central concern in Second Language Acquisition (SLA). Addressing fossilized errors requires a theoretical lens that explains not only how learners acquire language but also how they respond to corrective feedback. This study draws primarily on the Noticing Hypothesis and Cognitive Theory, which together provide a strong foundation for evaluating both the effectiveness and pedagogical quality of AI-generated linguistic feedback.

The Noticing Hypothesis, proposed by Schmidt and further developed by Gass and Ellis, posits that for linguistic input to become "intake" and contribute to language development, learners must consciously *notice* specific features in that input (Szcześniak, 2024). In the context of fossilized errors, such noticing becomes critical: learners must become aware of the discrepancy between their current interlanguage and the target form. Corrective feedback—especially when it is clear, salient, and timely—serves as a trigger for this noticing process (Barrot, 2023). If feedback merely supplies the correct form without drawing attention to the error, learners may fail to recognize the gap (Siow Chin et al., 2021), and fossilized patterns are likely to persist. Therefore, in evaluating ChatGPT's feedback, it is essential to assess not only whether corrections are made, but whether they promote awareness and facilitate learner noticing.

Complementing this, Cognitive Theory frames second language acquisition as a mental process involving attention, memory, and problem-solving (Mayer, 2024). Learners actively construct their linguistic knowledge through analysis, synthesis, and evaluation. Within this framework, feedback plays a regulatory role—it informs learners about the accuracy of their hypotheses, reinforces correct forms, and supports error elimination (Lipnevich & Smith, 2022). For feedback to be effective, however, it must be cognitively manageable and meaningful. Overly dense or vague corrections may lead to cognitive overload or disengagement, while targeted and clearly explained feedback can support deeper processing and long-term retention (Park & Ahn, 2022). Thus, from a cognitive perspective, evaluating ChatGPT's feedback involves assessing its clarity, scope, and potential to support metalinguistic reflection.

These two theoretical perspectives directly inform the study's research questions. RQ1 focuses on the extent to which ChatGPT accurately identifies and corrects fossilized errors—addressing the prerequisite condition for any effective intervention. RQ2 examines whether the feedback aligns with pedagogical principles derived from the Noticing Hypothesis and Cognitive Theory, including the promotion of learner awareness, meaningful engagement with error, and cognitive appropriateness. Together, these theories provide a focused and coherent lens for analyzing ChatGPT's role in addressing fossilization in L2 academic writing.

Previous Studies, Research Gap, and Novelty

The rapid integration of Artificial Intelligence (AI) into second language learning environments has led to a growing body of research investigating the efficacy of AI tools in supporting English as a Foreign Language (EFL) learners' writing development. Previous studies have explored the potential of automated writing evaluation (AWE) systems such as *Grammarly*, *Criterion*, and *ETS's e-rater*, highlighting their benefits in enhancing grammatical accuracy, vocabulary usage, and overall coherence in student writing (Fan, 2023; Khasawneh, 2024; Suryanto et al., 2024; Wei et al., 2023; Yildiz & Kuru Gonen, 2024). These tools, while effective in addressing surface-level errors, have often been critiqued for their limited feedback quality and lack of pedagogical depth—focusing predominantly on correction without explanation or meaningful learner engagement.

With the emergence of Large Language Models (LLMs) like ChatGPT, more recent research has begun to examine the affordances of AI-powered tools that can provide interactive, context-aware, and personalized feedback. Studies have shown that ChatGPT can assist learners in generating coherent text, revising drafts, and receiving immediate linguistic support (Javier & Moorhouse, 2024; Xiao & Zhi, 2023). Some have also compared its performance to that of traditional AWE systems (Amin et al., 2024; Bucol & Sangkawong, 2025; Fei et al., 2024). Moreover, preliminary investigations into the use of LLMs for feedback generation suggest a promising capacity to scaffold learner autonomy and metalinguistic reflection, especially in informal digital learning contexts.

However, despite this growing interest, a significant research gap remains regarding ChatGPT's effectiveness in handling fossilized grammatical errors—those persistent and often ingrained inaccuracies that resist conventional correction. Existing studies have not yet isolated fossilized errors as a specific focus, nor have they systematically examined how well ChatGPT identifies and corrects these deep-seated issues. Furthermore, few studies have critically evaluated the pedagogical quality of the feedback generated by ChatGPT in light of Second Language Acquisition (SLA) theories, particularly the Noticing Hypothesis and Cognitive Theory, which emphasize the need for feedback to trigger awareness, promote hypothesis-testing, and be cognitively manageable.

Additionally, while ChatGPT is often praised for its ability to provide grammatically correct outputs, there is limited empirical work on whether its feedback truly supports the internalization of correct forms, especially in advanced EFL writing contexts where

fossilized structures are common. Most existing research focuses on general error correction or on improving writing fluency, with less attention paid to error explanation, learner noticing, and feedback transparency, which are central to fostering lasting linguistic development.

This study seeks to address these gaps by offering a linguistic and pedagogical evaluation of ChatGPT's feedback on fossilized grammatical errors in EFL learners' academic writing. Its novelty lies in (1) its focus on fossilization as a distinct phenomenon within SLA, (2) its use of a controlled hypothetical corpus containing targeted fossilized error types commonly observed in EFL contexts (e.g., verb tense, articles, prepositions, subject-verb agreement), and (3) its theoretical framing through the Noticing Hypothesis and Cognitive Theory, allowing for a nuanced assessment of how well AI-generated feedback supports learner awareness and cognitive processing. By integrating insights from both SLA theory and AI capabilities, this research contributes a deeper understanding of how LLMs like ChatGPT can be pedagogically optimized to support advanced second language writing instruction.

METHODS

Research Design

This study employed a convergent parallel mixed-methods design (Creswell & Creswell, 2023), integrating both quantitative and qualitative approaches to comprehensively evaluate ChatGPT's handling of fossilized grammatical errors in EFL academic writing. In this design, quantitative and qualitative data were collected and analyzed concurrently but independently, with the aim of triangulating findings to provide a richer understanding of both performance accuracy and pedagogical quality. This research recognizes the dual focus of the research, namely the extent to which ChatGPT corrects fossilized errors (RQ1) and the way its feedback aligns with SLA-informed pedagogical principles (RQ2), required both numerical evaluation and interpretive insight. The quantitative strand measured grammatical error correction (GEC) performance using accuracy, precision, recall, and F-score, while the qualitative strand applied thematic content analysis to assess the depth, clarity, and instructional value of ChatGPT's feedback. The integration of these strands in the interpretation phase allowed the study to move beyond surface-level performance metrics and offer pedagogically meaningful conclusions about ChatGPT's role in second language writing instruction.

The quantitative component of the study evaluated ChatGPT's grammatical error correction (GEC) performance through established metrics, including accuracy, precision, recall, and F-score (Lin, 2024). These were computed by comparing ChatGPT's corrections against expert-annotated gold standard revisions. This analysis provided objective insights into the model's ability to detect and correct fossilized errors. The qualitative component consisted of an in-depth content analysis of ChatGPT's feedback. This analysis focused on the linguistic and pedagogical qualities of the corrections, including clarity, relevance, consistency across error types, and alignment with principles derived from the Noticing Hypothesis and Cognitive Theory. While

primarily qualitative, this phase also incorporated quantitative elements—such as coding frequency and interrater agreement—to assess consistency and enhance reliability.

The mixed-methods design was selected to address both "to what extent" (RQ1, quantitative) and "how" (RQ2, qualitative) dimensions of ChatGPT's performance. By combining statistical rigor with pedagogical analysis, this approach enabled a more holistic understanding of the tool's potential and limitations in correcting fossilized errors—a challenge known to resist superficial correction and require cognitively supportive feedback. The integration of methods thus ensured that both technical accuracy and instructional quality were evaluated in a balanced and meaningful way.

Objects of the Research

The object of this study was a carefully constructed hypothetical corpus of academic writing samples designed to simulate authentic English as a Foreign Language (EFL) learner output. The corpus was developed with the specific aim of embedding fossilized grammatical errors—errors that persist despite extensive exposure to the target language and formal instruction. By using a hypothetical corpus rather than real student writing, the study was able to exercise precise control over both the types and distribution of errors, ensuring direct alignment with the research focus and minimizing extraneous variables.

The corpus consisted of 500 unique sentences, embedded in short academic-style paragraphs representing typical segments of essay writing, such as introductions, body paragraphs, and conclusions. Each sentence contained at least one, and often multiple, instances of fossilized grammatical errors, carefully selected based on patterns consistently reported in Second Language Acquisition (SLA) literature and EFL pedagogy. The targeted error categories included article misuse (e.g., *"The student submitted a assignment late"* instead of *"the assignment"*), verb tense errors (e.g., *"Yesterday, she go to the library"* instead of *"went"*), and prepositional errors (e.g., *"She is interested on linguistics"* instead of *"in"*). Additionally, the corpus incorporated issues related to voice (e.g., incorrect use of passive or active constructions such as *"The experiment performed well"* instead of *"was performed"*), as well as persistent errors involving subject-verb agreement, word form confusions (e.g., *"scientific"* vs. *"scientifically"*), and lexical misselection due to L1 transfer or overgeneralization.

Each sentence was crafted to reflect the linguistic profile of intermediate to advanced EFL learners, thus preserving ecological validity while retaining the experimental control necessary for systematic analysis. The surrounding paragraph context provided sufficient syntactic and semantic cues for ChatGPT to generate contextually appropriate corrections and feedback, while also enabling human evaluators to assess the appropriateness of the output more reliably. The decision to develop a corpus of 500 sentences was methodologically driven: it was large enough to support quantitative analysis (e.g., accuracy, precision, recall across grammatical categories) and sufficiently rich for qualitative investigation of the pedagogical features of ChatGPT's feedback. Although no actual learner demographic data were involved, the linguistic patterns reflected in the

corpus were designed to approximate the challenges experienced by a diverse population of EFL learners from various L1 backgrounds. In sum, this tailored corpus served as a robust and replicable foundation for evaluating ChatGPT's effectiveness in addressing one of the most persistent challenges in L2 writing: fossilized grammatical errors.

Data Collection

Data collection in this study followed a systematic and replicable process, centered on the submission of the hypothetical EFL academic writing corpus to ChatGPT and the documentation of its responses. The version of ChatGPT used was GPT-4, accessed via the public API to ensure controlled and consistent interaction parameters throughout the evaluation. At the core of this process was the structured dataset of 500 sentences, each containing one or more fossilized grammatical errors. For each sentence, a "gold standard" correction was prepared by two independent expert linguists with specialized backgrounds in Second Language Acquisition and EFL pedagogy. Where discrepancies between expert annotations arose, they were resolved through discussion and consensus to maintain the validity of the benchmark data.

To ensure pedagogically oriented responses from ChatGPT, a standardized prompt was carefully developed through iterative prompt engineering. The prompt instructed ChatGPT to act as an "academic writing tutor" and to "identify and correct all grammatical errors in the provided text" while also "explaining the reasons for each correction in a clear, concise, and pedagogically appropriate manner, suitable for an advanced EFL learner." This formulation was designed to elicit both error correction (for RQ1) and explanatory feedback (for RQ2), encouraging responses that aligned with SLA-informed principles of feedback effectiveness.

Each erroneous sentence was submitted individually to ChatGPT using an automated script that preserved uniformity in prompt delivery and ensured the systematic recording of outputs. For every item, three components were captured: the original erroneous sentence, ChatGPT's corrected version, and the associated explanatory feedback. All responses were stored in a structured database to facilitate later analysis. Following the initial data collection, a preliminary review of the responses was conducted to screen for technical anomalies such as malformed outputs, incomplete corrections, or non-responses. Any such issues were either corrected through re-submission or excluded from the dataset if unresolved. This multi-layered approach ensured that the data collected were both reliable and relevant to the study's analytical goals..

Data Analysis

To answer RQ1, the quantitative analysis focused on comparing ChatGPT's corrected outputs with expert-annotated "gold standard" corrections across the 500 constructed sentences. Each response from ChatGPT was evaluated using standard grammatical error correction (GEC) metrics, including accuracy, precision, recall, and F-score (Foody, 2023; Mahmoud et al., 2023). Accuracy measured the proportion of correct classifications (i.e., whether ChatGPT accurately identified an error or correctly left an already accurate

structure unchanged). Precision assessed how many of ChatGPT's proposed corrections were correct, reflecting its ability to avoid false positives—instances where it introduced unnecessary changes. Recall captured the proportion of actual errors that ChatGPT correctly identified and corrected, indicating its sensitivity to genuine fossilized errors. The analysis relied on the $F_{0.5}$-score, a weighted harmonic mean of precision and recall that gives greater emphasis to precision, which is often prioritized in GEC tasks to ensure that corrections are both accurate and contextually appropriate.

To facilitate this comparison, an error annotation scheme modeled after frameworks such as ERRANT was employed, enabling the alignment of edits between the original sentence, ChatGPT's output, and the human-corrected version (McDowell, 2023; Qin et al., 2023). This allowed for fine-grained categorization of grammatical error types (e.g., verb tense errors, article misuse, preposition choice), labeled using standardized tags such as *R:VERB:TENSE* or *M:DET*. The comparison of edits was carried out using phrase-based alignment techniques supported by widely used GEC evaluation metrics, including the M2 scorer . Custom scripts developed in Python were used to automate the analysis and generate aggregate performance statistics across the dataset.

To address RQ2, a qualitative content analysis was conducted on ChatGPT's explanatory feedback accompanying each correction. The goal was to assess the pedagogical quality, depth, and consistency of the feedback in light of SLA-informed criteria, particularly from the perspectives of the Noticing Hypothesis and Cognitive Theory. A coding framework was developed based on principles of effective written corrective feedback in L2 writing. Key dimensions of analysis included the clarity and explicitness of explanations, the presence of metalinguistic information, the specificity and completeness of responses, and the pedagogical tone conveyed. Additionally, the analysis considered whether the feedback encouraged learner reflection or self-correction (as opposed to simply supplying the correct form), and whether the explanations were consistent across similar grammatical structures. Special attention was also given to how well the feedback addressed issues associated with fossilization, such as L1 transfer and cognitive entrenchment.

Following the coding process, a thematic analysis was conducted to identify recurring patterns and instructional features in the feedback. Multiple readings of the data allowed for the generation of initial codes, which were subsequently organized into broader themes. These included patterns such as "rule-based explanation," "generic or vague feedback," "lack of contextual sensitivity," and "feedback that promotes metalinguistic awareness." Representative excerpts were selected to illustrate each theme and support interpretive claims about ChatGPT's pedagogical strengths and limitations. In addition to this qualitative examination, a quantitative dimension of consistency was also analyzed. This involved tallying the frequency of specific feedback strategies across the corpus to determine whether ChatGPT provided uniform explanations for recurring error types. For example, consistency in explaining article usage or subject-verb agreement was assessed by calculating how often similar rules were articulated for structurally equivalent errors. By combining these analytical approaches, the study aimed to move beyond superficial

assessments of grammatical correction and instead provide a deeper evaluation of how AI-generated feedback might facilitate or hinder learner engagement, noticing, and long-term development in academic writing contexts.

FINDINGS AND DISCUSSION

The findings of this study are presented in two subsections, corresponding to the two research questions. The first subsection details the manifestations of pragmatic failure in EFL learners' emails, while the second explores the extent to which AI grammar tools identified and addressed these failures.

Accuracy of Error Identification and Correction

The quantitative analysis of ChatGPT's grammatical error correction (GEC) performance revealed considerable variation in accuracy across different types of fossilized errors commonly found in EFL academic writing. While ChatGPT demonstrated strong competence in handling structurally defined and rule-governed grammatical errors, its performance declined when dealing with more nuanced, context-dependent, or stylistically embedded issues—characteristics that frequently underpin fossilization in second language acquisition. To provide a general overview of ChatGPT's performance across error types, the table below summarizes its accuracy in correcting various fossilized grammatical structures

Table 1
ChatGPT's Correction Accuracy by Error Type

Error Category	Accuracy (%)	Performance Level	Comments
Subject-Verb Agreement	100%	High	Fully accurate in identifying agreement, even in complex structures
Singular/Plural Nouns	96%	High	Strong lexical sensitivity to count/mass noun distinctions
Word Form	96%	High	Consistently corrects derivational and part-of-speech errors
Verb Forms, Voice, Conditionals, Word Order	92%	High	Effective with conjugations, transformations, and linear syntax
Articles	89%	Moderate	Occasional inconsistency in nuanced contexts
Modal Verbs	81%	Moderate	Corrects basic usage; struggles with modality nuance
Verb Tense	76%	Moderate	Difficulty with aspect, sequence, and contextually grounded usage
Sentence Structure	52%	Low	Often fails to resolve structural ambiguity or rephrase convoluted syntax
Non-idiomatic Expressions	46%	Low	Struggles to recognize or revise unnatural phrasing typical of fossilization
Connectors	38%	Low	Misuse or omission of discourse markers; lacks pragmatic sensitivity

Unclear Messages	36%	Low	Fails to disambiguate meaning or resolve vague references

ChatGPT achieved high accuracy—defined here as over 90%—in correcting several grammatical categories typically regarded as rule-based and syntactically constrained. Notably, the model reached 100% accuracy in subject-verb agreement, effectively identifying mismatches even within complex sentence structures. For instance, in the sentence *"The data, which was collected over several months, were analyzed carefully,"* ChatGPT correctly identified and adjusted both the relative clause and the main clause agreement, producing the corrected version: *"The data, which were collected over several months, was analyzed carefully."* This reflects a robust internalization of grammatical agreement rules, which aligns with principles from the Noticing Hypothesis—suggesting that consistent and salient corrective input supports learner awareness of structural patterns.

Similarly, ChatGPT demonstrated 96% accuracy in correcting singular/plural noun errors. A typical example involved the replacement of *"Many researches have shown the importance of this method"* with *"Many studies have shown the importance of this method,"* indicating strong sensitivity to morphological and lexical norms in academic English. This pattern extended to word form errors, where ChatGPT also achieved 96% accuracy, reliably correcting adverbial and adjectival forms when inappropriate derivations were used (e.g., identifying the need for *"systematically"* rather than *"systematic"* in an adverbial context).

Other categories—such as verb forms, word order, passive constructions, conditionals, and lexical choice—also demonstrated high performance, each with approximately 92% correction accuracy. For example, in the sentence *"He has went to the conference last week,"* ChatGPT accurately produced *"He went to the conference last week,"* correcting the misuse of the past participle within a past time reference. These findings suggest that the model excels in identifying errors where correction rules are clearly codified and less reliant on discourse-level interpretation.

Moderate Accuracy in Context-Sensitive Structures

In grammatical domains that typically exhibit greater context dependency and are often sites of fossilization, ChatGPT's performance was more moderate. For article usage, an overall accuracy of 89% was observed. While this figure is relatively high, it nonetheless reflects occasional miscorrections or omissions, particularly in cases involving abstract nouns or context-sensitive definiteness. For instance, the model successfully corrected *"Importance of education is well-known"* to *"The importance of education is well-known,"* yet in more ambiguous cases, its ability to determine the appropriate use of definite or indefinite articles was inconsistent. This reflects known challenges in article acquisition for EFL learners, often tied to L1 transfer and subtle semantic distinctions, which may not be fully captured by statistical or rule-based models.

Performance in modal verbs (81%) and verb tense usage (76%) also reflected moderate accuracy. In the sentence *"The study found that the climate is changing rapidly over the past decade,"* ChatGPT correctly revised the tense to *"has been changing,"* aligning temporal reference with past actions extending into the present. Nevertheless, errors involving aspectual nuance, conditionality, or temporal sequencing in complex discourse were not consistently corrected. These limitations highlight the model's challenges with finer grammatical distinctions that are both cognitively demanding for learners and typically resistant to correction—hallmarks of fossilized constructions.

Low Accuracy in Discourse-Level and Idiomatic Errors

ChatGPT's performance declined significantly in correcting errors that require deeper interpretation of discourse, pragmatics, or idiomatic appropriateness. In categories such as unclear or ambiguous messages (36%), connective usage (38%), sentence structure (52%), and non-idiomatic expressions (46%), the model struggled to produce accurate and contextually appropriate corrections.

A typical example involved the sentence *"The committee decided to implement the new policy after much discussion, which was difficult."* ChatGPT's revision—*"The committee decided to implement the new policy after much difficult discussion"*—failed to fully resolve the underlying ambiguity of the referent "which was difficult," indicating the model's limited capacity for semantic disambiguation and pragmatic refinement. Similarly, in revising *"The research makes a good point about the issue,"* ChatGPT offered *"The research highlights a valid argument about the issue."* While the revision demonstrates some improvement in formality, the model inconsistently identifies or corrects non-native-like phrasing that, while grammatical, deviates from idiomatic academic English. These issues often result from L1 interference or insufficient exposure to naturalistic input and are emblematic of advanced-stage fossilization.

Pedagogical Appropriateness and Consistency of Linguistic Feedback

The qualitative analysis of ChatGPT's linguistic feedback on fossilized grammatical errors revealed a multifaceted picture. While the model is capable of generating grammatically accurate and metalinguistically informed corrections, its pedagogical appropriateness—particularly for fostering long-term learning and interlanguage restructuring—remains inconsistent. The feedback ranged from highly informative and explicit to overly generic or insufficiently explanatory, particularly for errors rooted in fossilized L2 habits. To provide a general overview, the following table summarizes ChatGPT's feedback performance across key pedagogical criteria.

Table 2
Evaluation of ChatGPT's Linguistic Feedback Across Pedagogical Dimensions

Pedagogical Dimension	Observed Strength	Observed Limitation
Clarity and Explicitness	Frequently explains basic rules clearly	Less consistent for complex or abstract grammatical issues
Metalinguistic Information	Provides grammar rules for well-defined errors	Often lacking in idiomatic or context-sensitive feedback
Specificity	High for verb forms, agreement, and determiners	Generic for articles, tenses, and collocations
Consistency	Consistent in recurring rule-based errors	Inconsistent for errors involving subtle distinctions or discourse context
Pedagogical Tone	Neutral, generally formal and non-judgmental	May be too mechanical; lacks adaptive tone based on learner profile
Encouragement of Self-Correction	Rare; mostly uses direct correction	Few prompts for reflection, noticing, or elicitation
Contextual Appropriateness	Adequate in local sentence corrections	Lacks understanding of learner background, level, and textual coherence
Idiomaticity and Register	Can revise toward academic tone	Struggles to explain why expressions are non-idiomatic or too informal

Comprehensiveness vs. Specificity

ChatGPT often provides comprehensive feedback, particularly on rule-based errors. For example, in correcting "He has a good knowledge of the subject," ChatGPT revised the sentence to "He has good knowledge of the subject" and offered a precise explanation regarding the uncountable nature of "knowledge." This form of feedback is clear, metalinguistically rich, and theoretically aligned with the Noticing Hypothesis, which emphasizes the learner's attention to gaps in their interlanguage system.

However, this comprehensiveness does not always translate into targeted specificity, especially for nuanced or fossilized errors. While ChatGPT tends to describe what was corrected, it often does not address why a learner might have made the error in the first place—such as due to L1 transfer or overgeneralization—limiting its effectiveness in guiding cognitive restructuring.

Consistency of Feedback

In domains with high grammatical accuracy, such as verb forms, ChatGPT's feedback is also consistently pedagogical. In the sentence "The researcher has went to the lab," the model not only corrected the verb to "went" but also clearly explained that "went" is the past form of "go" and that "has gone" would be the correct past participle. This kind of consistent rule reinforcement is beneficial for habit formation and error pattern recognition.

However, this consistency diminished in lower-performing categories. For example, article errors sometimes received detailed grammatical explanations (e.g., about countability), but similar errors in other contexts were corrected with vague comments such as "corrected article usage." This variability in explanatory depth can confuse learners and undercuts the scaffolding necessary for persistent error elimination.

Pedagogical Appropriateness and Explainability

One of the major shortcomings of ChatGPT's feedback is its limited explainability, particularly in revisions involving idiomatic language or register appropriateness. For instance, when revising "The findings give a good picture of the situation" to "The findings provide a clear overview of the situation," ChatGPT described the change as an improvement in academic tone. While the correction is stylistically valid, the lack of deeper explanation—such as insights into lexical collocation norms or typical academic phrasing—limits the learner's ability to internalize idiomatic usage patterns. This reflects a broader pattern in which stylistic corrections lack semantic justification or register-based reasoning, both of which are essential to developing near-native competence.

Learner-Specific Context and Over-Revision

ChatGPT also operates without awareness of the learner's proficiency level, curriculum, or writing goals, leading at times to overcorrections. In one case, it revised "While the cost was high, the benefits were significant, so we decided to proceed" to "Despite the high cost, the benefits were significant; thus, the decision to proceed was made." While grammatically sound, this syntactic overhaul might be too advanced or opaque for a learner at the intermediate level, offering little pedagogical scaffolding. Such over-revisions may not support language awareness development, and instead promote dependence on full replacement rather than internal hypothesis refinement.

Encouragement of Self-Correction and Noticing

Crucially, ChatGPT rarely uses strategies that promote learner noticing, self-monitoring, or hypothesis testing—processes that are central to overcoming fossilization. Most feedback is in the form of direct correction with explanation, rather than elicitation, recasts, or metacognitive prompts. For example, in changing "The discussion focused on the implications for the policy" to "The discussion focused on the implications of the policy," the feedback was correct but minimal: "Changed 'for' to 'of'. 'Implications of' is the correct prepositional phrase here." This correction lacks a prompt for the learner to

evaluate their own usage or reflect on the interlanguage gap, diminishing opportunities for cognitive engagement.

ChatGPT demonstrates significant strengths in generating clear, accurate, and grammatically oriented feedback for structurally simple fossilized errors. It performs particularly well in high-frequency grammatical domains where the correction rules are well-established and align with its large-scale training data. Its consistency in rule-based explanations supports learner noticing and potentially reinforces correct form recognition. However, when feedback involves context-sensitive, idiomatic, or discourse-level revisions, the model's pedagogical appropriateness declines. Explanations tend to become generic, less metalinguistically rich, or inflexible, and rarely incorporate learner-responsive features such as developmental appropriateness, scaffolding, or personalized support. Moreover, the model's default use of direct correction limits its utility in fostering metacognitive strategies, learner autonomy, and deep restructuring of interlanguage systems—particularly necessary for addressing fossilization. Thus, while ChatGPT can serve as a useful feedback generator, it lacks the interactional nuance and pedagogical intentionality required to fully replace human-mediated corrective feedback in SLA contexts.

DISCUSSION

The findings of this study provide a nuanced understanding of ChatGPT's capabilities and limitations in addressing fossilized grammatical errors in L2 academic writing. Quantitative results revealed that ChatGPT performs with high accuracy in correcting rule-governed grammatical structures such as subject-verb agreement, noun number, and basic verb forms. These results are consistent with prior research indicating that large language models (LLMs) excel at identifying surface-level grammatical errors that follow predictable rules (Alsaweed & Aljebreen, 2024). Compared to other AWE tools such as Grammarly or Criterion, which primarily rely on fixed rule-based algorithms, ChatGPT demonstrates more dynamic and context-aware feedback generation. For instance, while Grammarly has been shown to effectively flag mechanical errors, it often lacks explanation depth and interactive guidance (Fan, 2023; Khasawneh, 2024; . In contrast, ChatGPT's feedback—although inconsistent in complexity—occasionally includes metalinguistic explanations and paraphrasing options that reflect a more dialogic form of assistance. Moreover, ChatGPT's accuracy across categories aligns with findings by Xiao and Zhi (2023), who emphasized the model's strength in correcting conventional grammatical issues but noted a drop in performance when feedback required semantic inference or deeper discourse awareness. Taken together, these comparisons suggest that ChatGPT may serve as a more conversational and flexible tool than traditional AWEs, particularly as a first-pass proofreading aid. For L2 learners, this could reduce overreliance on instructors for low-level corrections and promote faster revision cycles—provided they are trained to interpret AI-generated feedback critically.

However, the analysis also uncovered significant performance drops in categories typically associated with fossilized errors, including article usage, verb tense consistency,

sentence structure, and non-idiomatic expressions. These errors are often deeply entrenched due to factors such as first language (L1) transfer, overgeneralization, and limited exposure to native-like input, making them resistant to simple corrective strategies. In these areas, ChatGPT's accuracy declined, suggesting that while it can provide grammatically acceptable alternatives, it often lacks the semantic and pragmatic sensitivity required to interpret more contextually or idiomatically appropriate forms (Dentella et al., 2024). For instance, article usage in English frequently depends on discourse-level meaning and shared background knowledge—subtleties that ChatGPT, operating primarily on statistical probabilities and sentence-level prompts, may not fully capture. Similarly, the model's limited ability to detect and revise non-idiomatic phrasing signals challenges in approximating naturalness and appropriateness in academic discourse.

The qualitative findings further illuminate the pedagogical implications of ChatGPT's feedback. While the model frequently offers comprehensive explanations, including metalinguistic rules and grammar terminology, the depth, consistency, and appropriateness of these explanations vary (Widyasari et al., 2024). For straightforward errors, the feedback is often detailed and clear. However, when addressing more complex or fossilized errors, the explanations are sometimes overly generic or missing altogether. This pattern suggests a tendency toward direct correction without cognitive scaffolding, which may limit learners' opportunity to process and internalize the feedback (Zhang & Zhang, 2023). According to the Noticing Hypothesis, effective feedback should not only correct the learner's output but also highlight the mismatch between the learner's interlanguage and the target form in a way that triggers reflection (Szcześniak, 2023; Szcześniak, 2024). Similarly, the Output Hypothesis emphasizes the importance of hypothesis testing through production and feedback. In this regard, ChatGPT's feedback often lacks the interactive quality necessary for fostering these deeper learning processes.

An additional concern is the model's inability to tailor feedback to individual learner profiles. Unlike human instructors, who adjust their comments based on learners' developmental stages, proficiency levels, and prior knowledge, ChatGPT provides standardized, context-free responses (Leon & Vidhani, 2023). This limitation may lead to either cognitive overload—where the learner is overwhelmed by dense or overly complex feedback—or under-challenging instruction, where more advanced learners receive insufficiently detailed guidance. The issue of over-revision, where learners accept ChatGPT's suggestions without critical evaluation, further highlights the importance of cultivating digital literacy in AI-supported learning environments (Joseph et al., 2024). In summary, ChatGPT demonstrates strong performance in correcting formal, rule-based errors but remains limited in its pedagogical effectiveness for addressing fossilized errors that require more nuanced, context-sensitive, and learner-centered intervention. While its error detection and correction capabilities are promising, especially for surface-level issues, ChatGPT cannot yet replicate the rich, adaptive, and explanatory feedback provided by experienced language instructors (Ouyang et al., 2024). The inherent "black-box" nature of LLMs continues to pose challenges for explainability, transparency, and trust in AI-mediated language learning.

The findings of this study carry several important pedagogical implications for EFL educators, curriculum designers, and educational policymakers regarding the integration of AI tools like ChatGPT into second language writing instruction. First, the evidence supports the use of ChatGPT as a complementary tool rather than a replacement for human feedback (Williyan et al., 2024). Its proficiency in correcting rule-based errors—such as those involving subject-verb agreement, pluralization, and basic verb forms—can help reduce teacher workload and allow instructors to focus on more cognitively demanding aspects of writing instruction, including organization, coherence, and the remediation of fossilized errors. This supports a hybrid feedback model, in which ChatGPT handles low-level corrections, while teachers provide personalized, formative guidance.

Second, the study highlights the need for explicit instruction in AI literacy. L2 learners must be equipped with the skills to critically evaluate and reflect upon AI-generated feedback (Ziqi et al., 2024). Since ChatGPT may offer generic, vague, or contextually inappropriate suggestions—particularly for complex grammar or stylistic issues—students should be taught to analyze the rationale behind corrections, compare them with their own understanding, and consult authoritative resources or instructors when necessary. This form of critical engagement encourages metacognitive awareness and strengthens learners' self-monitoring strategies.

Third, for curriculum development, the findings suggest that AI tools can support accuracy-focused tasks, but cannot substitute explicit teaching of complex grammar prone to fossilization. Curriculum designers should continue to prioritize targeted instructional interventions in areas such as article usage, verb aspect, and idiomaticity. Activities such as corpus-based analysis, contrastive analysis, and consciousness-raising tasks remain crucial (Li et al., 2025). These activities not only raise learners' grammatical awareness but also encourage deeper cognitive engagement with language patterns, thereby complementing the quick-feedback capabilities of AI tools.

Fourth, the findings underscore the importance of explainability in grammar correction systems. Developers of LLM-based educational tools should prioritize designing models that not only correct grammatical forms but also provide clear, context-sensitive, and pedagogically sound explanations (Mannekote et al., 2024). Features such as scaffolded feedback, learner interaction prompts, and metalinguistic cues can enhance the instructional quality of AI feedback and support more durable language learning. Incorporating adaptive feedback strategies based on learner profiles would further improve the relevance and impact of these systems.

Finally, from a policy perspective, the responsible and effective integration of AI into language education requires clear guidelines. Policymakers should promote ethical and pedagogically informed AI use, ensuring that these tools are positioned as supplementary aids rather than full replacements for human instruction. Academic integrity concerns must also be addressed, including potential overreliance on AI tools in student writing. Investment in teacher training and research into AI-mediated SLA is essential to ensure

that technological innovation genuinely enhances, rather than undermines, the language acquisition process.

CONCLUSION

This study provided a comprehensive evaluation of ChatGPT's effectiveness in addressing fossilized grammatical errors in EFL academic writing, with a particular focus on both the linguistic accuracy of its corrections and the pedagogical quality of its feedback. The findings revealed that ChatGPT performs impressively in correcting rule-based, surface-level errors such as subject-verb agreement, pluralization, and basic verb forms. However, its performance declined significantly in error categories that are more context-sensitive and indicative of fossilization, including articles, verb tenses, sentence structure, and non-idiomatic expressions. While ChatGPT often delivered detailed and seemingly comprehensive feedback, it frequently lacked the depth, contextual nuance, and learner-adaptive explanations necessary for fostering genuine understanding. Instead, the model tended to offer direct corrections with limited pedagogical scaffolding, thereby constraining its utility for long-term language development.

The contribution of this study lies in its targeted focus on a specific and persistent challenge in second language acquisition—fossilized grammatical errors. By utilizing a systematically designed corpus to elicit these errors, the research goes beyond general GEC evaluations to shed light on the pedagogical shortcomings of current AI tools in addressing deeply ingrained interlanguage issues. The findings underscore that while large language models can support certain aspects of writing instruction, particularly through automation and efficiency, they fall short in offering the kind of explanatory, individualized feedback that learners need to restructure their linguistic systems. The role of explainable AI, therefore, becomes crucial—not merely to correct errors, but to engage learners in a process of reflection and hypothesis testing that is central to overcoming fossilization.

Nonetheless, the study is not without limitations. While the use of a systematically constructed, hypothetical corpus allowed for precise control over error types and ensured consistent benchmarking, it also introduces ecological limitations. Real student writing often contains unpredictable variation, overlapping errors, idiosyncratic phrasing, and contextual dependencies that are difficult to replicate in a controlled setting. These characteristics could influence how ChatGPT performs in authentic contexts, potentially resulting in different patterns of correction accuracy or feedback clarity. As such, the findings, while robust within the defined experimental parameters, may not fully generalize to real-world instructional scenarios. Future research should incorporate authentic learner texts to validate and extend these results, enabling a deeper understanding of ChatGPT's pedagogical utility across diverse learner populations and writing contexts.

The results are tied to a specific version of ChatGPT and limited to grammatical fossilization in writing, excluding other domains such as lexical, phonological, or pragmatic fossilization. Moreover, the absence of actual learner interaction precludes

insight into how such feedback is received, internalized, or acted upon. Future research should address these gaps through longitudinal studies involving real learners, comparative evaluations of different AI systems, and the development of models designed specifically for pedagogical explainability. Ultimately, the integration of AI in language education holds great promise, but its most impactful use will emerge from a balanced human-AI partnership, one that leverages the efficiency of technology while preserving the depth, empathy, and adaptiveness of expert human instruction.

REFERENCES

Albelihi, H. H. M., & Al-Ahdal, A. (2024). Overcoming error fossilization in academic writing: Strategies for Saudi EFL learners to move beyond the plateau. *Asian-Pacific Journal of Second and Foreign Language Education*, *9*(1), 75. https://doi.org/10.1186/s40862-024-00303-y

Ali, S. (2024). A multidimensional analysis of academic writing: A comparative study of Saudi and British university students' writing. *World Journal of English Language*, *14*(2), 452. https://doi.org/10.5430/wjel.v14n2p452

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

Amin, M. M., Mao, R., Cambria, E., & Schuller, B. W. (2024). A wide evaluation of ChatGPT on affective computing tasks. *IEEE Transactions on Affective Computing*, *15*(4), 2204–2212. https://doi.org/10.1109/TAFFC.2024.3419593

Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, *36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Bucol, J. L., & Sangkawong, N. (2025). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, *62*(3), 867–882. https://doi.org/10.1080/14703297.2024.2363901

Canagarajah, S. (2024). Decolonizing academic writing pedagogies for multilingual students. *TESOL Quarterly*, *58*(1), 280–306. https://doi.org/10.1002/tesq.3231

Creswell, J. W., & Creswell, J. D. (2023). Research Design: Qualitative, quantitative and mixed methods approaches. In *SAGE Publications,Inc.: Vol. Sixth Edit* (Issue 1). SAGE Publications. https://medium.com/@arifwicaksanaa/pengertian-use-case-a7e576e1b6bf

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1), 28083. https://doi.org/10.1038/s41598-024-79531-8

Fan, N. (2023). Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study. *Sage Open*, *13*(2).

https://doi.org/10.1177/21582440231181296

Fei, X., Tang, Y., Zhang, J., Zhou, Z., Yamamoto, I., & Zhang, Y. (2024). Evaluating cognitive performance: Traditional methods vs. ChatGPT. *DIGITAL HEALTH*, *10*. https://doi.org/10.1177/20552076241264639

Foody, G. M. (2023). Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PLOS ONE*, *18*(10), e0291908. https://doi.org/10.1371/journal.pone.0291908

Javier, D. R. C., & Moorhouse, B. L. (2024). Developing secondary school English language learners' productive and critical use of ChatGPT. *TESOL Journal*, *15*(2). https://doi.org/10.1002/tesj.755

Joseph, G. V., P, A., Thomas M, A., Jose, D., V Roy, T., & Prasad, M. P. (2024). Impact of digital literacy, use of AI tools and peer collaboration on AI assisted learning-perceptions of the university students. *Digital Education Review*, *45*, 43–49. https://doi.org/10.1344/der.2024.45.43-49

Khasawneh, M. A. S. (2024). Investigating the impact of automated instruments used for assessing the writing skill: Perspectives of language e-learners. *Research Journal in Advanced Humanities*, *5*(2). https://doi.org/10.58256/4fd2qt78

Kormos, J. (2023). The role of cognitive factors in second language writing and writing to learn a second language. *Studies in Second Language Acquisition*, *45*(3), 622–646. https://doi.org/10.1017/S0272263122000481

Kos, Ž., & Mažgon, J. (2025). The challenges of using large language models: Balancing traditional learning methods with new technologies in the pedagogy of sociology. *Education Sciences*, *15*(2), 191. https://doi.org/10.3390/educsci15020191

Leon, A. J., & Vidhani, D. (2023). ChatGPT needs a chemistry tutor too. *Journal of Chemical Education*, *100*(10), 3859–3865. https://doi.org/10.1021/acs.jchemed.3c00288

Li, D., Noordin, N., Ismail, L., & Cao, D. (2025). A systematic review of corpus-based instruction in EFL classroom. *Heliyon*, *11*(2), e42016. https://doi.org/10.1016/j.heliyon.2025.e42016

Lin, S. (2024). Evaluating LLMs' grammatical error correction performance in learner Chinese. *PLOS ONE*, *19*(10), e0312881. https://doi.org/10.1371/journal.pone.0312881

Lipnevich, A. A., & Smith, J. K. (2022). Student – Feedback interaction model: Revised. *Studies in Educational Evaluation*, *75*, 101208. https://doi.org/10.1016/j.stueduc.2022.101208

Long, M. (2015). *Second language acquisition and task-based language teaching* (Vol. 11, Issue 1). Wiley Blackwell.

Mahmoud, Z., Li, C., Zappatore, M., Solyman, A., Alfatemi, A., Ibrahim, A. O., & Abdelmaboud, A. (2023). Semi-supervised learning and bidirectional decoding for effective grammar correction in low-resource scenarios. *PeerJ Computer Science*, *9*, e1639. https://doi.org/10.7717/peerj-cs.1639

Mannekote, A., Davies, A., Pinto, J. D., Zhang, S., Olds, D., Schroeder, N. L., Lehman, B., Zapata-Rivera, D., & Zhai, C. (2024). Large language models for whole-learner support: opportunities and challenges. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1460364

Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review*, *36*(1), 8. https://doi.org/10.1007/s10648-023-09842-1

McDowell, L. (2023). Japanese scientists' English for research publication purposes. *Journal of English for Research Publication Purposes*, *4*(2), 109–139. https://doi.org/10.1075/jerpp.22007.mcd

Ouyang, F., Guo, M., Zhang, N., Bai, X., & Jiao, P. (2024). Comparing the effects of instructor manual feedback and ChatGPT intelligent feedback on collaborative programming in China's higher education. *IEEE Transactions on Learning Technologies*, *17*, 2173–2185. https://doi.org/10.1109/TLT.2024.3486749

Park, J.-H., & Ahn, S. (2022). L2 learners' cognitive and behavioral engagement with written corrective feedback. *English Teaching*, *77*(3), 133–152. https://doi.org/10.15858/engtea.77.3.202209.133

Qin, Y., Luo, Y., & Zhai, Y. (2023). French error type annotation for dictation: A platform with automatic error type annotation for French dictation exercises. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1075932

Siow Chin, C., Pillai, S., & Zainuddin, S. Z. (2021). Recasts, prompts and noticing: A comparative study. *Studies in English Language and Education*, *8*(2), 416–441. https://doi.org/10.24815/siele.v8i2.18546

Suryanto, S., Habiburrahim, H., Akmal, S., Zainuddin, Z., Safrul, M. S., & Hanani, F. (2024). Scrutinizing the impacts of Grammarly application on students' writing performance and perception. *Jurnal Ilmiah Peuradeun*, *12*(2), 465. https://doi.org/10.26811/peuradeun.v12i2.1235

Szcześniak, K. (2023). There is more to learning words than meets the conscious eye. *Roczniki Humanistyczne*, *71*(10sp), 139–154. https://doi.org/10.18290/rh237110sp-7

Szcześniak, K. (2024). The noticing hypothesis and formulaic language. Learnability of non-salient language forms. *Acta Psychologica*, *248*, 104372. https://doi.org/10.1016/j.actpsy.2024.104372

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on

second language writing skills of Chinese EFL learners: a randomized controlled trial. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1249991

Widyasari, R., Zhang, T., Bouraffa, A., Maalej, W., & Lo, D. (2024). Explaining explanations: An empirical study of explanations in code reviews. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3708518

Williyan, A., Fitriati, S. W., Pratama, H., & Sakhiyya, Z. (2024). AI as co-creator: Exploring Indonesian EFL teachers' collaboration with AI in content development. *Teaching English With Technology*, *24*(2), 5–21. https://doi.org/10.56297/vaca6841/LRDX3699/RZOH5366

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages*, *8*(3), 212. https://doi.org/10.3390/languages8030212

Yildiz, H., & Kuru Gonen, S. I. (2024). Automated writing evaluation system for feedback in the digital world: An online learning opportunity for English as a foreign language students. *Turkish Online Journal of Distance Education*, *25*(3), 183–206. https://doi.org/10.17718/tojde.1169727

Zhang, X., & Zhang, R. (2023). Feedback, response, and learner development: A sociocultural approach to corrective feedback in second language writing. *Sage Open*, *13*(1). https://doi.org/10.1177/21582440231157680

Ziqi, C., Xinhua, Z., Qi, L., & Wei, W. (2024). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning*, 1–21. https://doi.org/10.1080/09588221.2024.2422478