AI-Generated Feedback in Secondary EFL Writing: A Scoping Review of an Emerging Evidence Base

Nwana, Edith Ngozi

Department of Art Education, Faculty of Education, University of Abuja, Abuja, Nigeria.
Authors Email: ngozinwana15@gmail.com; tel. number: 07035515915

Direct Research Journal of Social Science and Educational Studies

Direct RESEARCH JOURNALS

Vol. 14(3), Pp. 1-23, August 2026,

Review Article
ISSN: 2449-0806
https://journals.directresearchpublisher.org/index.php/drjsses; https://www.ajol.info/index.php/drjsses

Received 18 June 2026, Accepted 25 July 2026 Published 9 August 2026

ABSTRACT

Generative artificial intelligence (GenAI) has become increasingly prominent in second language writing instruction, yet empirical research on AI-generated written corrective feedback among secondary-school English as a Foreign Language (EFL) and English as a Second Language (ESL) learners remains limited. This scoping review mapped the available empirical evidence on AI-generated feedback and AI-generated text in secondary and middle-school EFL/ESL writing. The review was guided by the Joanna Briggs Institute scoping review framework and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR). The review employed a Population–Concept–Context framework, focusing on secondary and middle-school EFL/ESL learners, AI-generated written corrective feedback and related automated writing support, and formal secondary or middle-school writing instruction. Searches were conducted iteratively in June 2026 using a general academic web-search tool, and potentially relevant sources were screened and verified against their original publications. A single reviewer conducted the search, screening, and verification. Two studies met the eligibility and verification criteria: Ekizoğlu and Demir (2025), involving 60 Turkish secondary-school students, and Woo et al. (2025), involving 59 Hong Kong secondary-school students across seven schools. No formal quality appraisal or risk-of-bias assessment was conducted. Data were charted according to country and educational context, sample characteristics, research design, AI intervention or tool, comparison condition, outcome measures, and principal findings, and synthesized descriptively in relation to three research questions concerning writing accuracy and revision, learner engagement, and methodological and contextual gaps. The findings showed that AI-generated feedback was associated with significantly greater gains in overall writing performance, particularly grammar, vocabulary, and coherence, in the Turkish study, while 87% of students in the experimental group reported increased confidence in their writing. In the Hong Kong study, composition length was the strongest predictor of writing quality, while the relationship between AI-generated text use and writing performance varied according to academic banding. Among lower-banded students, attempts to revise AI-generated text were sometimes associated with lower composition scores. Overall, the evidence indicates that the benefits of AI-generated feedback at the secondary level are uneven and appear to vary according to learner proficiency, academic context, and patterns of engagement with AI-generated material. The review identifies a small and methodologically heterogeneous evidence base concentrated in only two national contexts and highlights the need for further research to establish the conditions under which AI-generated feedback may support secondary-school EFL/ESL writing development.

***Keywords**: Generative artificial intelligence, written corrective feedback, Secondary EFL, ESL writing, Scoping review, ChatGPT*

Citation: Nwana.E. N. (2026). AI-Generated Feedback in Secondary EFL Writing: A Scoping Review of an Emerging Evidence Base. *Direct Research Journal of Social Science and Educational Studies*. Vol. *14*(3): 1-23. https://doi.org/10.4314/drjsses.v14i3.1

INTRODUCTION

Generative artificial intelligence has moved from a peripheral curiosity to a routine presence in second language writing instruction within a remarkably short span of time. Since OpenAI released ChatGPT in late November 2022, researchers across applied linguistics and educational technology have produced a rapidly

growing body of work examining how large language models support, complicate, or transform the teaching of written English to learners of English as a Foreign Language (EFL) and English as a Second Language (ESL). One systematic review found that within just eighteen months of the tool's release, seventy empirical studies had already examined ChatGPT's applications in ESL and EFL education, a volume of output that itself signals how quickly this research area has expanded (Lo et al., 2024). Among the many pedagogical functions these tools now perform, the provision of written corrective feedback has attracted particular scholarly attention, in part because feedback has long occupied a contested but central position in second language writing theory, and in part because ChatGPT's capacity to generate fluent, immediate, and seemingly individualized commentary appears to offer a direct technological answer to one of the oldest practical constraints in language teaching, namely the limited time teachers have to respond to every learner's draft.

What this expanding literature has not yet done is treat the secondary school context as a distinct unit of analysis. Lo et al. (2024) reported that of the seventy studies included in their review, only two were conducted in a K to 12 setting, with the remaining majority situated in higher education, and the authors explicitly called for further research focused on early childhood, primary, and secondary settings, concluding that findings drawn almost entirely from university populations could not be assumed to generalize downward to younger learners. That gap, identified in 2024, has not remained empty. Within roughly the following eighteen months, a small but identifiable cluster of studies began to examine secondary level EFL learners specifically. A subsequent scoping review by Crosthwaite and Sun (2025), synthesizing fifty-one studies on generative AI and second language written feedback, mapped the broader field comprehensively but did not isolate secondary or K to 12 learners as a distinct analytic category, leaving the scattered secondary level studies it included undifferentiated from the surrounding higher education literature. Ekizoğlu and Demir (2025), for instance, conducted a quasi-experimental study with sixty Turkish secondary school students, finding that those who received feedback from an AI writing assistant showed significantly greater gains in writing performance, particularly in grammar, vocabulary, and coherence, than students who received only teacher feedback. Comparable studies have since examined Hong Kong secondary students' interaction with AI generated text during composition, among others. These studies have not yet been gathered together and read as a body. Each currently exists as a comparatively isolated data point within a literature still organized around the higher education default that Lo et al. (2024) identified. This absence matters for reasons beyond bibliographic completeness. Secondary school learners differ from university students in ways that plausibly affect how they engage with AI generated feedback. They are typically less proficient in English, less practiced in independent self-regulated learning, and in most jurisdictions are minors, which raises distinct ethical and pedagogical questions about exposure to generative AI systems that are largely absent from research on adult university populations. Ekizoğlu and Demir (2025) addressed this directly in their own study, obtaining informed consent from both students and parents and securing institutional ethics board approval before proceeding, a procedural step that reflects the heightened duty of care required when adolescent learners are the research population. A finding that holds for an adult undergraduate using ChatGPT to revise an argumentative essay cannot simply be assumed to hold for a fourteen- or fifteen-year-old EFL learner receiving AI generated feedback in a secondary classroom. Until the secondary level evidence is read as its own body of work, claims about the generalizability of AI feedback research will continue to rest on an evidentiary base that the field has already acknowledged is too narrow.

The present review addresses this gap directly. Given the small size and methodological heterogeneity of the available secondary level evidence, the review adopts scoping review methodology, which is designed for mapping emerging bodies of literature rather than for pooling effect estimates across studies. It draws together the empirical literature on AI generated and teacher generated written corrective feedback in secondary and middle school EFL and ESL writing classrooms, synthesizing what this still emerging body of evidence indicates about effectiveness, learner engagement, and the practical conditions under which AI feedback succeeds or breaks down at this level of education. The review is guided by two theoretical lenses long established in second language writing research. Written corrective feedback theory, shaped substantially by the long running debate between Ferris and Truscott over whether and how corrective feedback contributes to language development, provides the substantive framework for evaluating what these studies report about accuracy and revision outcomes. The Noticing Hypothesis, developed by Schmidt, offers an explanatory mechanism for why the source of feedback, whether generated by a machine or delivered by a human teacher, might plausibly affect whether learners attend to and internalize the corrections they receive. Three research questions organize the synthesis that follows. First, what does the available secondary level evidence indicate about the comparative effectiveness of AI generated and teacher generated feedback on writing accuracy and revision behavior. Second, how do adolescent EFL and ESL learners engage with AI generated feedback, both behaviorally and affectively, when it is introduced into secondary classroom settings? Third, what methodological and contextual gaps remain in this body of research, and what do they suggest for the design of future studies. In addressing these questions,

the review aims to give secondary level evidence the dedicated treatment it has not yet received, and to offer EFL and ESL educators working with adolescent learners a clearer account of what can, and cannot, currently be claimed about AI generated feedback at their level of practice.

Literature Review and Theoretical Framework

Written Corrective Feedback Theory

The theoretical point of departure for the present review is the longstanding debate concerning the efficacy and pedagogical value of written corrective feedback (WCF) in second-language writing. The debate predates the emergence of generative artificial intelligence by several decades and is principally associated with the contrasting positions advanced by Truscott (1996) and Ferris (1999). Truscott questioned the efficacy of grammar correction in second-language writing instruction, arguing that correction may consume considerable instructional resources without necessarily producing durable improvements in learners' underlying linguistic competence. Ferris (1999), in contrast, challenged this position, contending that the evidence available at the time did not warrant such a categorical rejection of corrective feedback and that appropriately implemented feedback could contribute to greater linguistic accuracy. The subsequent exchange between these positions established a foundational theoretical problem in second-language writing research: whether the provision of corrective information translates into sustained changes in learners' developing linguistic systems or merely facilitates local and potentially transient improvements in written products.

The significance of this debate for the present review extends beyond its historical importance. It provides a conceptual framework for distinguishing between feedback provision, learner engagement with feedback, revision, and durable language development. This distinction is particularly important in the context of generative AI, where the capacity to produce immediate, extensive, and linguistically sophisticated feedback can easily be mistaken for evidence of learning. The pedagogical value of AI-generated feedback cannot therefore be inferred solely from the quantity or apparent accuracy of corrections supplied. Rather, its effectiveness must be considered in relation to whether learners notice the identified problem, understand the rationale underlying the suggested modification, evaluate its appropriateness, and subsequently transfer the acquired knowledge to new writing contexts. In this respect, the central concern raised by Truscott regarding correction without substantive learning remains theoretically relevant, even though the technological mechanism through which correction is delivered has changed substantially. At the same time, contemporary evidence cautions against interpreting the Truscott position as a general indictment of corrective feedback. Brown et al. (2026), in a Bayesian meta-analysis of 52 primary studies, reported robust evidence that WCF produces moderate effects that can persist across short-, medium-, and long-term measurement periods. Their findings are particularly consequential because they provide a more methodologically sophisticated basis for evaluating the durability of WCF than earlier aggregations of the literature. Importantly, the meta-analysis also indicated broadly comparable effects across direct, indirect, and metalinguistic forms of feedback. These findings lend empirical weight to the more qualified position associated with Ferris: corrective feedback, when appropriately implemented, can contribute to the development of L2 writing accuracy and should not be dismissed on the grounds that correction is necessarily transient or pedagogically ineffective.

Nevertheless, the contemporary literature also demonstrates that the effectiveness of corrective feedback cannot be reduced to a simple dichotomy between effective and ineffective correction. Rahimi et al. (2025), for example, found that automated written corrective feedback (AWCF) delivered within an activity-theory-informed instructional environment resulted in superior performance among EFL learners compared with a non-electronic feedback condition, particularly in overall writing performance, task achievement, and grammatical range and accuracy. However, the intervention did not produce significant differences in coherence and cohesion or lexical development. This pattern is theoretically important because it suggests that the effects of automated feedback may be dimension-specific rather than globally transferable across all components of writing proficiency. Moreover, the learners' behavioral, cognitive, and affective engagement with AWCF indicates that the educational consequences of automated feedback are likely to depend not simply on what the system provides, but also on how learners interact with and respond to that feedback.

The importance of learner engagement is further reinforced by research examining individual differences in responses to WCF. Ni and Xu (2025), drawing on control-value theory, demonstrated that learners' emotional responses to corrective feedback are not necessarily uniform. Although English proficiency had a relatively limited influence on the eight emotional responses examined, learning motivation was found to play a substantially more important role, with highly motivated learners experiencing both positive and negative emotions more intensely than their less motivated counterparts. This finding complicates any account of feedback effectiveness that treats learners as passive recipients of corrective information. It suggests instead that affective and motivational processes may shape how feedback is experienced and potentially acted upon. For the present review, this is particularly significant because AI-generated feedback may alter not only the source and speed of correction but also learners' perceptions of

feedback, autonomy, responsibility, and engagement during revision. The temporal dimension of corrective feedback also warrants theoretical consideration. Cheng and Zhang (2025) found that both synchronous and asynchronous computer-mediated WCF significantly improved linguistic accuracy, although neither condition produced corresponding improvements in syntactic complexity and fluency. More importantly, synchronous feedback was more effective than asynchronous feedback in promoting accuracy. These findings suggest that the timing and interactional configuration of feedback may influence its effects, rather than feedback operating as a context-independent intervention. Such evidence is relevant to generative AI because AI systems can provide feedback in highly immediate and interactive forms, potentially altering the temporal relationship between writing, feedback, revision, and learner reflection. However, immediacy should not automatically be equated with deeper learning. If rapid feedback reduces the need for learners to diagnose linguistic problems themselves, the efficiency of correction could theoretically coexist with limited development of independent error-recognition and self-regulatory capacities.

The emergence of generative AI introduces a further dimension to this theoretical debate by changing not merely the delivery mechanism of WCF but also the nature, scalability, and responsiveness of automated corrective intervention. Wu et al. (2025) reported that generative AI models, particularly GPT-4 when supported by task-specific prompting, could outperform commercial automated writing tools in error detection and correction, while participant feedback suggested potential improvements in writing quality and confidence and reductions in teacher workload. However, the reported concern regarding occasional overcorrection is theoretically consequential. Greater technological sophistication does not eliminate the fundamental pedagogical problem identified within the WCF literature: feedback must be appropriately interpreted and used by learners if it is to contribute to development. A system can therefore become increasingly accurate at identifying and correcting linguistic forms without necessarily ensuring that the learner develops the capacity to recognize, explain, and independently reproduce those forms. These studies suggest that the contemporary evidence provides stronger support for a conditional rather than categorical interpretation of WCF effectiveness. The evidence does not sustain Truscott's strongest implication that corrective feedback is inherently incapable of producing durable linguistic development; the recent Bayesian meta-analysis provides evidence of moderate and durable effects. At the same time, neither does the evidence justify a straightforwardly Ferris-like assumption that the provision of correction will reliably generate improvement across learners, writing dimensions, and instructional conditions. Rather, current research points towards a more complex explanatory model in which the effects of corrective feedback is mediated by the characteristics of the feedback itself, the timing and mode of delivery, the linguistic dimension targeted, learners' cognitive and affective engagement, and the instructional environment within which feedback is received and acted upon. This synthesis provides an important theoretical basis for examining AI-generated feedback in secondary-school EFL writing. Generative AI potentially amplifies several of the affordances associated with automated WCF, including immediacy, individualization, scalability, and repeated opportunities for revision. Yet these affordances do not, in themselves, establish that AI-generated feedback promotes durable writing development. The critical theoretical question is therefore not simply whether generative AI can identify and correct linguistic errors, but whether its feedback facilitates the cognitive, affective, and behavioral processes through which learners transform correction into transferable linguistic knowledge. This distinction is especially important when considering secondary-school learners, whose developmental characteristics, levels of writing expertise, metacognitive resources, and capacity for critically evaluating automated feedback may differ considerably from those of the tertiary learners who dominate much of the existing evidence.

Accordingly, the present review adopts a position that moves beyond the binary "feedback works" versus "feedback does not work" formulation that has historically characterized the WCF debate. Instead, it conceptualizes AI-generated corrective feedback as a pedagogical intervention whose effects are likely to be conditional, mediated, and differentiated across learners and writing outcomes. The enduring contribution of the Ferris–Truscott debate lies precisely in this unresolved question of mechanism: correction becomes educationally consequential not merely when it is supplied, but when learners engage with it in ways that promote noticing, understanding, revision, and subsequent application. Contemporary evidence on automated and generative-AI-mediated WCF therefore does not make the earlier theoretical debate obsolete. Rather, it provides a technologically transformed context in which the central question can be reformulated with greater precision: under what conditions, for which learners, and through which mechanisms do AI-generated corrective feedback contribute to sustained development in second-language writing? It is this theoretically and empirically unresolved question that provides the foundation for the present review.

The Noticing Hypothesis

A second theoretical lens underpinning the present review is Schmidt's Noticing Hypothesis, which provides a cognitive account of how exposure to linguistic input may become available for subsequent language development. Developed independently of the written corrective feedback debate, the hypothesis has

nevertheless, become closely associated with research on corrective feedback because it offers a plausible explanation of the cognitive processes through which learners may convert corrective information into linguistic knowledge. Schmidt (1990) proposed that conscious attention to linguistic features plays a critical role in second-language acquisition, distinguishing between the broader input to which learners are exposed and the portion of that input that becomes intake through processes of attention and awareness. Within this formulation, noticing constitutes a theoretically important interface between exposure and learning: linguistic information may be present in the input, but its pedagogical significance depends, at least in part, on whether learners attend to and process the relevant form. The relevance of the Noticing Hypothesis to AI-generated corrective feedback lies in its capacity to explain why feedback provision does not necessarily equate to feedback learning. Corrective feedback may identify a linguistic discrepancy, but the identification of that discrepancy does not guarantee that the learner attends to it, understands its significance, or incorporates the relevant form into subsequent language use. From this perspective, the effectiveness of AI-generated feedback depends not only upon the technical accuracy of the correction but also upon the extent to which the feedback renders the relevant linguistic feature sufficiently salient for learner attention. The distinction is particularly consequential in generative-AI environments, where feedback can be produced instantaneously, repeatedly, and at considerable volume. Greater availability of corrective information may increase opportunities for noticing, but it may equally overwhelm learners or encourage uncritical acceptance of suggested revisions if the feedback is not accompanied by appropriate cognitive engagement.

Recent scholarship provides both theoretical refinement and empirical support for applying the Noticing Hypothesis to technology-mediated language learning. Barrot (2023), for example, reported gains in L2 writing accuracy following the use of Grammarly and attributed these gains, in part, to the tool's capacity to facilitate noticing, provide adaptive metalinguistic explanations, and support self-directed learning. This finding is significant for the present review because it suggests that automated feedback can potentially perform more than a mechanical error-detection function. When feedback draws learners' attention to discrepancies and provides information that enables them to understand those discrepancies, automated systems may create conditions conducive to the cognitive processing envisaged by the Noticing Hypothesis. The theoretical implication is that the pedagogical value of automated feedback resides not simply in its capacity to supply the correct form but in its potential to make the relationship between the learner's production and the target language sufficiently salient to stimulate further processing. This interpretation is consistent with the broader conceptualization of noticing advanced by Bogdanov (2026), who argues that the construct has frequently been used ambiguously in second-language acquisition research, with noticing variously conflated with awareness, selective attention, discrepancy detection, depth of processing, intake, and learning itself. This conceptual instability is important for the present review because evidence that learners have encountered or attended to AI-generated feedback should not automatically be interpreted as evidence that acquisition has occurred. Bogdanov's process-oriented reconceptualization positions noticing as a transient and relational boundary within ongoing communicative activity rather than as a directly observable mental state. Such a formulation encourages greater caution when interpreting AI-mediated feedback studies: observed revision, correction acceptance, or immediate improvement in a written product may provide evidence of behavioral response to feedback, but these outcomes do not necessarily establish that noticing has resulted in durable linguistic development. The distinction between noticing and learning is further reinforced by emerging evidence that challenges a strong interpretation of Schmidt's original claim. Szcześniak et al. (2026), for instance, reported evidence that novel lexical sequences could be retained following incidental exposure even when learners' attention was directed towards meaning rather than linguistic form. Their findings suggest that some aspects of language learning may occur without deliberate conscious attention to the target linguistic feature, thereby challenging the stronger proposition that conscious noticing is invariably necessary for acquisition. This evidence does not render the Noticing Hypothesis irrelevant; rather, it indicates that its explanatory power may be more appropriately understood as identifying one important route through which learning can occur rather than establishing an exclusive prerequisite for all forms of acquisition.

For the present review, this qualification is particularly important because AI-generated feedback may operate through both explicit and less deliberate forms of learning, depending upon the learner, task, feedback design, and linguistic feature involved. Shahini (2025) similarly identifies the enduring influence of the Noticing Hypothesis in language-learning theory while acknowledging criticisms concerning its emphasis on conscious learning and the difficulty of empirically measuring noticing. The distinction between intentional and incidental learning is therefore central to the present theoretical positioning. AI-generated feedback may deliberately direct learners' attention towards specific errors through explicit correction, explanations, highlighting, or prompts for revision. At the same time, repeated exposure to corrected forms may potentially facilitate learning processes that are not fully accessible to conscious awareness. The theoretical implication is that AI-mediated feedback should not be evaluated solely according to whether learners can explicitly report having

noticed a correction; rather, attention should also be given to subsequent processing, retention, transfer, and independent use of the relevant linguistic forms. Evidence from EFL research further illustrates the potential interaction between noticing and instructional intervention. Elkhettm and Habiba (2026) reported significant improvement in grammatical performance following instruction informed by the Noticing Hypothesis and related input-oriented principles. Although the study involved a very small university sample and therefore provides limited grounds for broad generalization, its findings are relevant in demonstrating the pedagogical possibility of deliberately directing learners' attention towards linguistic forms. More importantly, the study's conclusion that intentional and incidental learning may complement rather than necessarily exclude one another supports a more flexible interpretation of noticing in the context of AI-mediated instruction. The relevance of proficiency-sensitive noticing becomes particularly apparent in emerging research on generative AI. Juwita et al. (2025), examining the use of Gemini among beginner language learners, reported a strong positive relationship between post-task noticing and delayed recall and found substantial differences in retention between A1 and A2 learners. The study therefore provides preliminary evidence that AI-generated language input can facilitate attention to linguistic features and that noticing may be associated with subsequent retention. However, the marked difference between proficiency groups is theoretically significant. It suggests that the effectiveness of AI-mediated noticing may depend partly upon learners' existing linguistic resources and their ability to process and interpret the language presented to them. Consequently, the assumption that the same AI-generated feedback will produce comparable learning effects across learners at different proficiency levels requires careful scrutiny. This proficiency dimension has direct implications for the secondary-school focus of the present review. Unlike the university populations that dominate much of the existing AI-mediated language-learning literature, secondary-school EFL learners may demonstrate substantial variation in linguistic proficiency, metalinguistic awareness, digital literacy, and capacity for independent evaluation of AI-generated feedback. The same corrective explanation may therefore be highly salient and comprehensible to one learner but linguistically inaccessible or cognitively overwhelming to another. AI-generated feedback may consequently facilitate noticing for some learners while producing superficial correction, selective attention, or even disengagement for others. The question of whether AI feedback promotes noticing must therefore be considered alongside the question of whose attention is being directed, towards what linguistic feature, under which conditions, and with what subsequent consequences for learning. The issue of immediacy provides a further point of theoretical convergence between the Noticing Hypothesis and AI-generated feedback. The immediacy of AI-mediated correction may increase the temporal proximity between learner production and corrective information, potentially making discrepancies more salient than feedback delivered after a considerable delay. In secondary-school contexts, this possibility has been invoked to explain why learners may identify and attend to errors that might otherwise remain unnoticed under conventional teacher-marked drafts. However, immediacy should not be treated as an independent guarantee of noticing or learning. The mere presentation of immediate feedback does not establish that learners have processed its linguistic significance, understood the explanation, or retained the relevant form. The theoretical value of immediacy therefore lies in its potential to create favorable conditions for noticing, rather than in its capacity to produce acquisition automatically. The literature suggests that the Noticing Hypothesis remains a valuable but necessarily qualified theoretical framework for analyzing AI-generated corrective feedback. Its principal contribution is to explain the cognitive pathway through which corrective information may become pedagogically meaningful: feedback must attract attention and be processed by the learner if it is to have the potential to influence subsequent language use. However, contemporary scholarship also cautions against equating noticing with acquisition or treating conscious attention as the sole mechanism of language development. Evidence concerning incidental learning, conceptual ambiguity surrounding the measurement of noticing, and differences associated with learner proficiency all suggest that the relationship between attention, feedback, and learning is more complex than a linear model of feedback → noticing → acquisition would imply.

Accordingly, the present review conceptualizes noticing not as a guaranteed or sufficient mechanism of AI-assisted language development but as a potentially important mediating process through which AI-generated feedback may influence writing development. This position allows the review to accommodate both evidence supporting the cognitive importance of attention and evidence questioning the necessity of conscious noticing for all forms of language learning. It also provides a theoretically defensible basis for examining why AI-generated feedback may produce different outcomes across secondary-school learners. The central analytical concern is therefore not simply whether AI feedback is noticed, but whether, and under what conditions, noticing develops into meaningful processing, revision, retention, and transferable language knowledge. This distinction is fundamental to the present review's broader argument that the educational effectiveness of generative AI cannot be inferred from its capacity to generate corrections alone; rather, effectiveness must be understood in relation to the cognitive and contextual processes through which learners engage with, interpret, and ultimately appropriate the feedback they receive.

The Empirical Landscape Prior to Generative AI

Before generative AI systems entered the writing classroom, written corrective feedback research had already established a substantial empirical base concerning automated and technology mediated feedback more broadly. Automated Writing Evaluation systems, exemplified by tools such as Criterion and Grammarly, had been studied for their capacity to identify surface level errors in grammar, spelling, and mechanics, generally with findings suggesting these tools could support measurable gains in accuracy when integrated into a structured writing process rather than used in isolation. Wei, Wang, and Dong (2023), in a randomized controlled trial involving 190 Chinese EFL learners, found that students who received Grammarly based Automated Writing Evaluation training showed superior performance across multiple dimensions of writing skill, including task achievement, coherence and cohesion, lexical range, and grammatical accuracy, when compared to a control group, with learners' initial proficiency levels also shown to meaningfully predict the size of these gains. This body of work matters for the present review because it establishes that automated feedback, even before the advent of generative AI, was not a uniformly novel intervention but rather an extension of an existing line of research into how learners respond to feedback that originates from a source other than a human teacher, and because the proficiency dependent pattern Wei et al. (2023) identified anticipates, in a higher education context, the proficiency moderated pattern this review later finds among secondary learners specifically.

The Emergence of Generative AI as a Feedback Source

The emergence of generative artificial intelligence (GenAI) has introduced a qualitatively different form of automated feedback into second-language writing instruction. Although automated writing evaluation and automated written corrective feedback systems preceded generative AI, earlier systems were typically constrained by rule-based or comparatively bounded computational procedures for identifying linguistic errors and generating corrective information. The release of ChatGPT in November 2022 marked a significant technological shift by making conversational, context-sensitive, and comparatively fluent feedback accessible to learners through natural-language interaction. This development has consequently expanded the conceptual boundaries of automated feedback, from the identification and correction of discrete linguistic errors towards more dialogic forms of explanation, revision support, and interaction around written texts. The research response to this technological development has been exceptionally rapid. Lo et al. (2024), in their synthesis of 70 empirical studies examining ChatGPT in ESL and EFL education during the first eighteen months following its release, found that writing constituted the most frequently investigated language domain, accounting for more than 40% of the studies reviewed. However, the distribution of this research was markedly uneven, with higher education representing the dominant context. This pattern is important because it indicates that the rapid expansion of GenAI research has not been accompanied by an equivalent expansion of evidence across educational levels. The knowledge base concerning AI-mediated writing feedback has consequently developed primarily around university learners, leaving substantial uncertainty concerning its applicability to younger EFL learners whose linguistic, cognitive, metacognitive, and self-regulatory capacities may differ considerably from those of tertiary students.

The higher-education concentration is further substantiated by more recent scholarship. Zhang et al. (2026), in a scoping review of 51 empirical studies published between 2023 and 2025, found that GenAI-driven feedback research remains fragmented despite growing empirical attention. Their synthesis identified considerable variation in students' perceptions and engagement with GenAI feedback and, importantly, reported mixed evidence concerning both immediate effects and longer-term developmental trajectories. The review also identified human–GenAI integration, learner-centered prompt engineering, and student characteristics as important directions for future research. These observations challenge any interpretation of the rapid growth of GenAI feedback research as evidence that the field has already established a stable understanding of its educational effectiveness. Rather, the expanding evidence base appears to have generated a more complex set of questions concerning how feedback is produced, how learners interact with it, and under what conditions such interaction results in meaningful learning.

The theoretical implications of this development are substantial. The affordances of GenAI—including immediacy, scalability, repeatability, conversational responsiveness, and the capacity to generate explanations tailored to particular prompts—potentially address some of the practical constraints associated with teacher-mediated written corrective feedback. Lindsay et al. (2025) argue that generative AI makes rich, constructive feedback more scalable and repeatable while potentially reducing the burden associated with routine feedback provision. However, they simultaneously emphasize that responsible implementation requires attention to inclusivity, contextual relevance, expert oversight, and the risk that automated systems may fail to accommodate learners whose needs are less easily captured by standardized computational processes. Thus, the technological capacity to make feedback abundant does not necessarily make that feedback pedagogically appropriate or developmentally equitable.

This distinction is particularly relevant to the relationship between AI-generated feedback and learner agency. Otaki (2023), examining perceptions of generative-AI and

human feedback among students and educators in higher education, highlighted the emotional, relational, and dialogic dimensions of feedback and argued for consideration of AI feedback alongside, rather than simply in place of, human feedback. Similarly, Jensen et al. (2026) found that chatbot feedback differed from supervisory feedback in important ways: chatbot interactions were more strongly task-focused and performance-oriented, whereas human feedback was more relational, contextual, and developmental. Their findings suggest that generative AI can provide an additional source of information and support learner agency, but does not eliminate the need for human engagement in feedback processes that seek to promote meaningful learning.

Evidence from hybrid feedback environments further supports this position. Saini et al. (2025) found complementary strengths in AI-generated and peer feedback: AI feedback was comparatively strong in structured, precise, and consistent evaluation, whereas peer feedback contributed greater opportunities for critical reflection and human-centered insight. Of particular relevance was the finding that the sequencing of feedback modalities influenced learning outcomes, with AI feedback potentially establishing a foundation for subsequent engagement with peer feedback. This suggests that the pedagogical question may not be whether AI should replace human feedback, but how different feedback sources can be deliberately orchestrated to combine technological consistency with the relational and reflective dimensions of human interaction. Sajadi et al. (2024) similarly demonstrated the potential of GenAI to enhance peer-feedback processes by transforming self- and peer-generated comments into more constructive and actionable feedback. Collectively, these findings support a hybrid rather than replacement-oriented conception of AI-mediated feedback.

The technological potential of GenAI must nevertheless be considered alongside concerns regarding reliability and accuracy. Wu et al. (2025) found that generative AI models, including GPT-4, could outperform commercial automated writing tools in error detection and correction when appropriate prompting strategies were employed. Their findings also indicated potential benefits for writing quality and learner confidence and suggested that GenAI could reduce teacher workload. However, the authors simultaneously identified concerns about occasional overcorrection. This qualification is theoretically important because it demonstrates that improvements in the technical sophistication of AI feedback do not remove the need for learners to evaluate the validity and relevance of the feedback they receive. In other words, the increased linguistic fluency of an AI-generated explanation should not be conflated with pedagogical infallibility. The importance of learner characteristics is also becoming increasingly evident. Juwita et al. (2025), examining Gemini-supported language learning, reported a strong positive relationship between post-task noticing and delayed recall and identified substantial differences in retention between A1 and A2 learners. Although the study involved university students and therefore cannot be directly generalized to secondary-school populations, the findings are theoretically relevant because they suggest that AI-mediated learning effects may be sensitive to learners' existing proficiency. The higher retention observed among A2 learners compared with A1 learners indicates that the effectiveness of AI-generated linguistic input may depend partly on whether learners possess sufficient linguistic resources to process and interpret the material presented to them. Such evidence reinforces the need to treat learner characteristics as potential moderators rather than assuming that AI-generated feedback operates uniformly across proficiency levels.

The Secondary-Level Evidence Base

The limitations of the higher-education evidence become particularly consequential when considering secondary-school EFL learners. The emerging secondary-level literature remains considerably smaller than the tertiary evidence base, but the studies available provide important grounds for questioning the assumption that the benefits of GenAI feedback are uniform across adolescent learners. Rather than simply reproducing the optimism evident in some university-based studies, secondary-school evidence increasingly points towards a more conditional relationship between AI use and writing development.

Ekizoğlu and Demir (2025), for example, investigated the effects of AI-assisted writing feedback among 60 Turkish secondary-school students using a quasi-experimental design. Over an eight-week intervention involving weekly writing tasks, the experimental group received feedback from an AI writing assistant, whereas the control group received traditional teacher feedback. The AI-feedback group demonstrated greater gains in overall writing performance, with particularly notable improvements in grammar, vocabulary, and coherence. The study also reported that 87% of students in the experimental group felt more confident about their writing after using the AI tool. The findings therefore provide preliminary evidence that AI-mediated feedback may support both selected dimensions of writing development and learners' affective perceptions of their writing competence. Nevertheless, the study's quasi-experimental design and specific Turkish secondary-school context warrant caution against treating these findings as universally generalizable. A substantially different picture emerges from Woo et al. (2025), whose study examined 59 Hong Kong secondary-school students across seven schools. Rather than comparing AI feedback with teacher feedback, the researchers examined the relationship between the incorporation of AI-generated text and the quality of students' compositions. Their findings indicated that the total

number of words in a composition, irrespective of whether those words were AI-generated or student-authored, was the strongest predictor of writing quality. However, the relationship between AI use and writing performance varied according to students' academic banding. Students in higher-banded schools appeared more capable of using AI-generated text strategically and of producing high-quality compositions without extensive dependence on AI. Among lower-banded students, heavier reliance on AI-generated text was associated with higher composition scores, but attempts to revise AI-generated text in their own words were sometimes associated with lower scores. This finding introduces an important qualification to the assumption that greater learner intervention with AI-generated material necessarily represents greater pedagogical benefit.

The contrast between Ekizoğlu and Demir (2025) and Woo et al. (2025) is theoretically productive rather than merely contradictory. The former provides evidence consistent with the proposition that AI-assisted feedback can enhance selected aspects of secondary-school writing performance and learner confidence, whereas the latter demonstrates that the relationship between AI use and writing quality may vary according to learners' existing academic positioning and their manner of engaging with AI-generated text. These findings suggest that the central issue is not whether AI is beneficial or detrimental in an absolute sense, but how learner characteristics and pedagogical conditions mediate the relationship between AI assistance and writing development.

This interpretation is consistent with the broader evidence concerning learner engagement and individual differences. Rahimi et al. (2025), for instance, found that learners receiving automated written corrective feedback electronically outperformed those receiving non-electronic WCF in overall writing performance, task achievement, and grammatical range and accuracy, but not in coherence and cohesion or lexical development. Their qualitative evidence further indicated behavioral, cognitive, and affective engagement with automated feedback. Such findings reinforce the proposition that AI-mediated feedback may influence particular dimensions of writing more readily than others and that learner engagement is an important component of the process through which feedback may become educationally productive. Similarly, Brown et al.'s (2026) Bayesian meta-analysis provides an important corrective to any interpretation that positions contemporary evidence as simply reproducing Truscott's skepticism. Across 52 primary studies, the authors found robust evidence for moderate and durable effects of WCF over short-, medium-, and long-term periods, with broadly similar effects across direct, indirect, and metalinguistic forms of feedback. However, the persistence of WCF effects at the general level does not eliminate the need to examine variation across learners, contexts, writing dimensions, and feedback conditions. The secondary-school findings are therefore important precisely because they allow the broader WCF evidence to be interrogated within a population whose developmental and educational characteristics may differ from those represented in much of the existing literature. The evidence concerning learner characteristics also extends beyond proficiency alone. Ni and Xu (2025) found that English learning motivation played a substantially greater role than English proficiency in shaping learners' emotional responses to WCF, with highly motivated learners experiencing both positive and negative emotions more intensely. This finding suggests that the consequences of corrective feedback cannot be understood exclusively through measurable writing outcomes. Learners' affective responses, perceptions, motivation, and engagement may form part of the mechanism through which feedback contributes to, or fails to contribute to, subsequent development. In secondary-school environments, where learners may vary considerably in motivation and readiness for independent use of AI, such individual differences may be especially consequential.

The timing of feedback provides another potentially important explanatory variable. Cheng and Zhang (2025) found that both synchronous and asynchronous computer-mediated WCF significantly improved linguistic accuracy, while synchronous feedback was more effective than asynchronous feedback. However, neither treatment produced corresponding improvements in syntactic complexity or fluency. This finding cautions against treating immediate feedback as synonymous with comprehensive writing development. AI systems may offer unprecedented immediacy, but rapid correction may primarily influence dimensions of accuracy without necessarily producing equivalent gains in broader aspects of writing competence. Thus, the temporal advantage of GenAI should be interpreted as a potential affordance whose educational value depends upon how learners process and apply the feedback.

The emerging secondary-school evidence provides a substantially more differentiated picture than a simple narrative of technological enhancement. Ekizoğlu and Demir (2025) demonstrate potential benefits for writing performance and learner confidence; Woo et al. (2025) demonstrate that AI-related writing outcomes vary across academically differentiated groups and modes of engagement; Rahimi et al. (2025) show that automated feedback may benefit some dimensions of writing more than others; and the broader WCF literature demonstrates that feedback can produce durable effects without implying that those effects are uniform across all learners or instructional circumstances. The implications for the present review are therefore twofold. First, the available evidence does not support a technologically deterministic assumption that access to generative AI will inherently improve secondary-school EFL writing. Second, neither does the evidence justify a wholesale rejection of AI-generated feedback on the grounds that automated correction cannot contribute to learning.

Instead, the literature points towards a conditional and interactional account of AI-mediated writing development, in which outcomes emerge from the interaction among feedback characteristics, learner proficiency, motivation, engagement, timing, task demands, instructional mediation, and the broader educational context. This position also provides an important bridge between the two theoretical perspectives developed earlier in this review. Written Corrective Feedback Theory establishes the question of whether corrective intervention contributes to durable linguistic development, while the Noticing Hypothesis provides a potential cognitive explanation for how corrective information may become available for learning. The secondary-school evidence suggests that neither framework should be applied mechanistically. AI-generated feedback may make errors more immediately visible and provide repeated opportunities for revision, but noticing does not guarantee acquisition, and correction does not guarantee durable development. Conversely, evidence of improvement following AI-mediated feedback indicates that automated correction can, under appropriate conditions, contribute to writing development. The theoretical and empirical challenge is therefore to identify the conditions under which these possibilities are realized.

Consequently, the present review positions secondary-school AI-generated feedback not as an inherently beneficial or harmful innovation, but as a mediated pedagogical intervention whose effects are contingent upon learner, feedback, task, and contextual factors. This positioning is necessary because the current evidence base remains relatively small, methodologically heterogeneous, and disproportionately informed by tertiary education. The emerging secondary-school literature is beginning to reveal patterns that cannot safely be inferred from university populations, particularly with respect to proficiency, academic banding, confidence, engagement, and the manner in which learners incorporate or revise AI-generated material. The central contribution of the present review is therefore to bring this fragmented evidence into a coherent analytical framework and to determine whether the apparent benefits of generative AI in EFL writing represent consistent developmental effects or context-dependent outcomes shaped by the characteristics of learners and the pedagogical environments in which AI is deployed.

Synthesis and Positioning of the Present Review

Taken collectively, the literature reviewed in this section provides a basis for advancing three interrelated propositions concerning the use of AI-generated feedback in secondary-school EFL writing. First, the theoretical foundations underpinning the interpretation of written corrective feedback remain both relevant and analytically productive in the context of generative artificial intelligence. In particular, the Written Corrective Feedback (WCF) literature and Schmidt's Noticing Hypothesis provide complementary explanatory lenses through which the pedagogical value of AI-generated feedback can be interrogated. WCF theory directs attention to the nature, explicitness, accuracy, and pedagogical function of corrective intervention, whereas the Noticing Hypothesis foregrounds the cognitive processes through which learners attend to linguistic discrepancies and potentially convert corrective information into developing interlanguage knowledge. The emergence of generative AI, therefore, does not necessarily render these theoretical perspectives obsolete. Rather, it introduces a technologically mediated dimension through which established questions concerning correction, noticing, learner engagement, uptake, and subsequent linguistic development can be reconsidered. The theoretical challenge is consequently not one of wholesale replacement, but of determining how established constructs operate when feedback is generated, personalized, and delivered through an artificial rather than exclusively human agent.

Second, although empirical scholarship on generative AI and AI-mediated feedback in EFL/ESL writing has expanded considerably, the evidential base remains unevenly distributed across educational levels and contexts. A substantial proportion of the available research has been undertaken within higher education, where learners are typically older, more autonomous, and potentially more capable of critically evaluating and strategically using AI-generated feedback. This contextual concentration is consequential because findings derived from university populations cannot automatically be generalized to secondary-school learners. Adolescents differ from university students not only in linguistic proficiency but also in metacognitive development, digital literacy, self-regulation, writing experience, and capacity to distinguish between linguistically plausible feedback and pedagogically appropriate feedback. Consequently, the apparent effectiveness of AI-generated feedback within higher education may partly reflect learner characteristics and contextual conditions that are less prevalent at secondary level. The predominance of higher-education evidence therefore constitutes not merely a demographic limitation in the existing literature but a substantive threat to the transferability of current conclusions to adolescent EFL writing contexts.

Third, the emerging secondary-school evidence complicates the assumption that increased access to AI-generated feedback necessarily produces improved writing outcomes. Although some studies indicate potential benefits in relation to revision, linguistic accuracy, feedback accessibility, and learner engagement, the effects appear to be conditional rather than universally positive. In particular, evidence that learners with lower levels of proficiency may respond differently to, or derive fewer benefits from, AI-generated feedback raises an important question concerning the interaction between technological affordances and

learner characteristics. The findings reported by Woo et al. (2025), for example, that revision of AI-generated text was in some instances associated with lower composition scores among lower-banded students, challenge a technologically deterministic interpretation of AI-assisted writing. Such evidence is particularly significant when considered alongside Truscott's longstanding concern that corrective intervention does not automatically translate into durable linguistic development. The central issue, therefore, is not simply whether AI can identify or correct errors, but whether learners possess the linguistic, cognitive, and metacognitive resources necessary to interpret, evaluate, and meaningfully incorporate the feedback provided.

This distinction is critical to the present review because it shifts the analytical focus from the availability of AI-generated feedback to the conditions under which such feedback becomes pedagogically productive. An AI system may provide technically accurate, immediate, and extensive feedback, yet the educational value of that feedback ultimately depends upon learner noticing, interpretation, evaluation, uptake, and subsequent application. In this respect, AI-generated feedback should not be conceptualized as an autonomous mechanism of writing development. Rather, its effects are likely to emerge from an interaction among the quality and characteristics of the feedback, learner proficiency, task demands, opportunities for revision, teacher mediation, and the learner's capacity to engage critically with machine-generated suggestions. This position also helps reconcile apparently contradictory findings in the literature: evidence of positive effects and evidence of limited or even adverse effects need not necessarily represent mutually exclusive conclusions if the effects of AI-mediated feedback are contingent upon learner and contextual characteristics.

The literature consequently exposes a significant conceptual and empirical tension. On one hand, the rapid development of generative AI has expanded the possibilities for immediate, individualized, and scalable written corrective feedback. On the other, the pedagogical assumption that more feedback, faster feedback, or technologically sophisticated feedback will necessarily produce greater improvement remains insufficiently substantiated, particularly among adolescent EFL learners. The Ferris–Truscott debate is therefore not simply displaced by generative AI; rather, AI-mediated feedback provides a new empirical environment in which the underlying disagreement concerning the efficacy, durability, and pedagogical consequences of corrective feedback can be revisited. From this perspective, the question is no longer reducible to whether corrective feedback "works", but must encompass for whom, under what conditions, through which mechanisms, and with what consequences for subsequent writing development.

It is within this unresolved terrain that the present review is positioned. Existing scholarship has generated valuable evidence concerning AI-assisted writing and feedback, but the secondary-school literature remains comparatively fragmented, methodologically heterogeneous, and theoretically under-integrated. Studies differ substantially in their conceptualization of feedback, outcome measures, learner populations, intervention designs, modes of AI use, and definitions of writing improvement, thereby making direct comparison and cumulative interpretation difficult. Moreover, the limited volume of secondary-school research has not yet been sufficiently synthesized to establish whether the patterns observed across individual studies constitute consistent effects, context-dependent tendencies, or methodological artefacts.

Accordingly, the present review does not approach the literature with the assumption that AI-generated feedback is either inherently beneficial or inherently detrimental to secondary-school EFL writing development. Instead, it adopts a critical and theoretically informed position that treats effectiveness as conditional, mediated, and potentially differentiated across learners. The review therefore seeks to synthesize the available evidence across secondary-school contexts, examine the extent to which findings converge or diverge, interrogate the methodological and theoretical explanations for such variation, and identify the learner-, feedback-, and context-related conditions that may account for differential outcomes. In doing so, the review seeks to move beyond the binary question of whether AI-generated feedback is effective and towards a more nuanced account of how, why, for whom, and under what pedagogical conditions AI-mediated corrective feedback may contribute to sustainable writing development among secondary-school EFL learners.

Such positioning is particularly important because the expansion of generative AI into educational practice is occurring more rapidly than the accumulation of developmentally appropriate empirical evidence. Without a systematic synthesis of the secondary-school literature, there is a risk that pedagogical recommendations will be extrapolated from university-based research and that the distinctive cognitive, linguistic, and educational circumstances of adolescent learners will remain insufficiently represented. The present review therefore addresses not simply a gap in the quantity of research, but a gap in the organization, interpretation, and theoretical integration of existing evidence. Its contribution lies in bringing together a dispersed body of secondary-school evidence within an established theoretical framework while critically examining the extent to which current findings support, qualify, or challenge prevailing assumptions about corrective feedback and AI-mediated writing development.

METHODOLOGY

Review Design

This study employed a scoping review design to map and

critically synthesize the emerging empirical evidence concerning the use of generative artificial intelligence (GenAI)-mediated feedback and AI-generated text in secondary-school English as a Foreign Language (EFL) and English as a Second Language (ESL) writing. The review was guided by the methodological principles of the Joanna Briggs Institute (JBI) scoping review framework and reported with reference to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR). A scoping review was selected because the literature addressing GenAI-mediated writing support among secondary-school EFL/ESL learners remains small, recent, and methodologically heterogeneous. The available studies differ in research design, learner characteristics, AI application, comparison conditions, writing tasks, and outcome measures. Consequently, the evidence base does not currently provide an adequate foundation for statistical pooling or meta-analysis. Instead, the purpose of the present review was to identify and characterize the available secondary-level evidence, examine reported effects and patterns of learner engagement, and identify methodological, geographical, theoretical, and contextual gaps that require further investigation.

The review was specifically motivated by the disproportionate concentration of AI-mediated EFL/ESL writing research in higher education. Previous reviews have identified a substantial imbalance between tertiary and pre-tertiary research, creating uncertainty about whether findings derived primarily from adult university learners can be transferred to adolescent learners. The present review therefore treated secondary-school EFL/ESL learners as a distinct population rather than as a subgroup embedded within the broader literature on AI-assisted language learning. The review was additionally informed by two established theoretical perspectives: Written Corrective Feedback (WCF) theory and Schmidt's Noticing Hypothesis. WCF theory provided the principal framework for interpreting questions concerning corrective feedback, writing accuracy, revision, and the longstanding Ferris–Truscott debate, while the Noticing Hypothesis provided a cognitive framework for examining how learners may attend to, process, and potentially incorporate AI-generated linguistic input into subsequent writing. These theoretical perspectives were used during interpretation and synthesis rather than as criteria for determining study eligibility.

Review Questions

The review was guided by three research questions developed from the identified literature gap and the theoretical orientation of the study:

RQ1: What does the available secondary-school evidence indicate about the effects of AI-generated feedback and AI-generated text on writing accuracy, writing quality, and revision behavior?

RQ2: How do secondary-school EFL/ESL learners engage with AI-generated feedback and AI-generated text, both behaviorally and affectively?

RQ3: What methodological, geographical, educational, and contextual gaps characterize the emerging evidence on AI-mediated writing support among secondary-school EFL/ESL learners?

The questions were deliberately formulated to accommodate the methodological diversity of the available studies. In particular, the review did not assume that every study would employ a direct AI-versus-teacher comparison. This was necessary because the included studies examined different forms of AI-mediated writing support: one directly compared AI-assisted feedback with teacher feedback, whereas the other investigated the relationship between AI-generated text use and composition quality.

Eligibility Framework

Eligibility was structured using a Population–Concept–Context (PCC) framework, which was considered more appropriate than an intervention-based framework because the objective was to map an emerging and heterogeneous body of evidence rather than estimate the comparative effectiveness of a single intervention.

The Population comprised secondary-school and middle-school EFL/ESL learners, generally corresponding to learners approximately 11–18 years of age and enrolled in formal schooling before tertiary education.

The Concept comprised AI-mediated writing support, with particular emphasis on AI-generated written corrective feedback, AI-generated text used in writing activities, and closely related automated writing-support applications. This broader conceptualization was necessary because the included literature did not operationalize AI support uniformly. While Ekizoğlu and Demir (2025) directly investigated AI-assisted writing feedback, Woo et al. (2025) examined students' incorporation and revision of AI-generated text in their compositions.

The Context was formal secondary- or middle-school EFL/ESL writing education. Studies involving tertiary learners were not considered eligible unless secondary-level findings were reported separately and could therefore be interpreted independently.

Search Strategy

The search strategy was developed around three interrelated conceptual domains. The first comprised terms associated with generative artificial intelligence and specific AI applications, including generative artificial intelligence, generative AI, GenAI, ChatGPT, GPT, AI

chatbot, and related terminology. The second comprised terms associated with written corrective feedback and automated writing support, including written corrective feedback, AI-generated feedback, automated written corrective feedback, automated writing feedback, AI-assisted writing, and AI-generated text. The third comprised terms associated with the target population and educational setting, including EFL, ESL, English as a Foreign Language, English as a Second Language, secondary school, middle school, high school, adolescent learners, and K–12. The search was conducted in June 2026. Searching proceeded iteratively rather than through a single fixed search expression. Initial exploratory searches addressed AI-generated feedback and AI-mediated writing in EFL/ESL education broadly. Following identification of the substantial concentration of studies involving university populations, subsequent searches were narrowed to secondary-school, middle-school, and K–12 contexts. This iterative strategy was adopted because the terminology surrounding GenAI-mediated writing support remains unstable and varies across studies. Restricting the search to a single technical term, such as "AI-generated written corrective feedback", could have excluded relevant studies that described substantially similar practices using alternative terminology.

Information Source and Search Limitation

The search was conducted using a general academic web-search environment capable of identifying scholarly publications. Potentially relevant records were subsequently traced to their original publications for verification.

The search did not constitute a comprehensive multi-database search of Web of Science Core Collection, Scopus, ERIC, or other specialist bibliographic databases. This limitation is explicitly acknowledged because it affects the comprehensiveness and reproducibility of the evidence base. The present review should consequently be interpreted as a scoping evidence map of the identifiable literature, rather than as an exhaustive claim that all published secondary-level studies were retrieved. This limitation is particularly important given the rapid development of the GenAI literature and the possibility that relevant studies may be indexed under different terminology, publication formats, or disciplinary databases. A future systematic extension should replicate the search across Web of Science Core Collection, Scopus, ERIC, and relevant education and applied-linguistics databases, supplemented by backward and forward citation searching.

Eligibility Criteria

Inclusion Criteria

A study was considered eligible when it satisfied the following criteria:

Population: The study involved secondary-school, middle-school, or high-school learners engaged in EFL or ESL education, generally corresponding to learners approximately 11–18 years of age.

Writing focus: The study addressed English writing, writing development, writing performance, writing revision, or written corrective feedback.

AI-related concept: The study investigated GenAI-generated feedback, AI-assisted writing feedback, AI-generated text used within writing activities, or a closely related AI-mediated writing-support practice.

Educational context: The study was situated within formal secondary- or middle-school education.

Empirical evidence: The study reported original empirical findings derived from quantitative, qualitative, or mixed-methods research.

Population-specific evidence: Where participants were drawn from multiple educational levels, secondary-school findings had to be identifiable separately from tertiary or adult learner findings.

Verifiability: The authorship, publication details, and original source of the study had to be sufficiently identifiable to permit independent verification.

Exclusion Criteria

Sources were excluded if they:

(i) Focused exclusively on university, tertiary, or adult learners;
(ii) Examined AI use in language education without a substantive writing or written-feedback component;
(iii) Addressed general educational applications of GenAI without relevance to EFL/ESL writing;
(iv) Were theoretical articles, opinion pieces, editorials, commentaries, or non-empirical publications;
(v) Combined secondary and tertiary populations without separately reporting secondary-level findings;
(vi) Lacked sufficient information to establish the nature of the AI-mediated writing activity;
(vii) Could not be traced to an identifiable original publication; or
(viii) Contained incomplete or unverifiable authorship information.

The final criterion was particularly important in the present review because the objective was not simply to identify potentially relevant claims but to construct an evidence base in which every included study could be independently traced and verified.

Study Identification, Screening, and Verification

The study-selection process consisted of identification, relevance screening, eligibility assessment, and source verification. Potentially relevant records were initially assessed against the population, concept, context, and writing-related criteria. Records that appeared potentially eligible were subjected to further examination of their original publication details and empirical content. A single reviewer conducted the search, screening, eligibility assessment, and verification procedures. This constitutes an important methodological limitation because independent duplicate screening would provide greater protection against selection errors and subjective interpretation of eligibility criteria. To minimize this limitation, predefined eligibility criteria were applied consistently, potentially relevant sources were retained until their eligibility could be established, and included studies were traced to original publications before being incorporated into the evidence synthesis. The single-reviewer procedure should therefore be regarded as a transparent limitation of the review rather than as a methodological substitute for independent screening. Future replications should employ at least two independent reviewers and document agreement and disagreement during title/abstract and full-text screening.

Source Verification Procedure

Because the review was concerned with a rapidly developing literature in which secondary reports, preprints, and incomplete bibliographic records may circulate alongside formal publications, a specific verification procedure was incorporated into the review.
Each potentially eligible source was traced to its original publication. Author names, publication year, publication venue, study characteristics, and the substantive relationship between the study and the review questions were checked against the source itself rather than being accepted solely on the basis of secondary citations or search-engine summaries. This procedure resulted in the exclusion of two potentially relevant candidate studies. One concerned engagement with generative-AI feedback among Chinese middle-school learners, while another concerned AI-generated and teacher-generated model texts among Chinese high-school learners. Although the available information suggested that both studies were relevant to the review, complete and reliable authorship information could not be independently verified using the search resources available during the review. They were consequently excluded from the formal evidence base. Their identification is nevertheless methodologically relevant because it suggests that the secondary-school evidence base may be somewhat larger than the studies formally charged in this review. These records should be prioritized for verification in future replications.

Final Evidence Base

Following application of the eligibility and verification criteria, two studies were included in the final synthesis: Ekizoğlu and Demir (2025) and Woo et al. (2025). Ekizoğlu and Demir (2025) investigated AI-assisted writing feedback among 60 Turkish secondary-school students using a quasi-experimental design over an eight-week intervention.

The study compared an AI-feedback group with a teacher-feedback group and examined changes in writing performance, including grammar, vocabulary, and coherence. Woo et al. (2025) examined 59 Hong Kong secondary-school students across seven schools and investigated the incorporation and revision of AI-generated text within student compositions. Unlike Ekizoğlu and Demir (2025), the study did not employ direct AI-feedback versus teacher-feedback comparison. Instead, it examined relationships between AI-generated text use, student-authored text, composition length, academic banding, and writing quality. The methodological and conceptual differences between these studies were retained during synthesis rather than treated as grounds for exclusion, because such heterogeneity is itself informative in an emerging research field.

Data Charting

Data were charted systematically from each included study using a structured extraction framework. The charting fields were designed to capture information relevant to the review questions and the theoretical interpretation of the evidence.

The following information was extracted:

Author(s) and publication year;
Country and geographical context;
Educational setting and school level;
Participant/sample size;
Learner age or age range where reported;
Research design;
Nature of the AI technology or application;
Form of AI-generated feedback or AI-generated text;
Comparison condition, where applicable;
Duration and structure of the intervention or study;
Writing task or activity;
Writing outcomes assessed;
Revision behavior;
Learner confidence and other affective responses;
Behavioral or engagement patterns;
Principal findings;
Relevant methodological characteristics; and
Limitations and implications reported by the original study.

The extracted information was organized around the three review questions and subsequently interpreted in relation to WCF theory and the Noticing Hypothesis.

Data Synthesis and Analytical Approach

Because only two studies met the final eligibility and verification criteria, and because they differed substantially in research design, intervention structure, comparison conditions, and outcome measurement, meta-analysis and quantitative effect-size aggregation were not undertaken. Instead, the review employed a descriptive and narrative thematic synthesis. The synthesis proceeded in three stages.

First, the characteristics of the included studies were mapped descriptively according to country, school context, sample size, research design, AI application, feedback arrangement, and outcomes.

Second, empirical findings were organized around the three research questions. The first analytical domain concerned writing accuracy, writing quality, and revision behavior. The second concerned behavioral and affective engagement with AI-generated feedback or text. The third concerned methodological and contextual gaps in the emerging secondary-school evidence base.

Third, the findings were interpreted comparatively and theoretically. Particular attention was given to areas of convergence and divergence between the two studies, including the apparent benefits of AI-mediated feedback, differences associated with learner academic banding and proficiency, the distinction between AI feedback and AI-generated text, and the extent to which improvement in a particular written product can reasonably be interpreted as evidence of underlying writing development. The synthesis therefore did not treat the two studies as statistically comparable interventions. Rather, it treated them as complementary sources of evidence that illuminate different dimensions of AI-mediated writing support at the secondary level. This approach is consistent with the central purpose of a scoping review: to identify patterns, conceptual relationships, evidence gaps, and areas requiring more definitive investigation.

Theoretical Interpretation of the Evidence

The empirical synthesis was interpreted through two complementary theoretical perspectives: Written Corrective Feedback Theory and Schmidt's Noticing Hypothesis. WCF theory was used to examine whether the evidence supports the proposition that corrective feedback can contribute to writing development and to reconsider the longstanding Ferris–Truscott debate in the context of generative AI. Particular attention was given to the distinction between immediate improvement in written products and evidence of durable, transferable linguistic development. The Noticing Hypothesis was used to consider the cognitive processes through which learners may attend to linguistic discrepancies identified through AI-generated feedback and subsequently process, revise, retain, or transfer those forms. However, the review did not equate behavioral engagement, revision, or correction acceptance automatically with noticing or acquisition. This distinction was considered particularly important given the evidence that AI-mediated writing outcomes may differ according to learner proficiency and academic context. The theoretical synthesis therefore treated noticing as a potential mediating process rather than a sufficient or guaranteed mechanism of learning. Similarly, corrective feedback was conceptualized as potentially beneficial but conditional upon learner characteristics, feedback quality, timing, engagement, instructional mediation, and opportunities for subsequent independent application.

Methodological Appraisal

A formal risk-of-bias or methodological quality assessment was not used to determine eligibility because the primary purpose of this review was to map an emerging evidence base rather than calculate a pooled estimate of intervention effectiveness. Nevertheless, methodological characteristics were considered critically during interpretation. The review examined differences in study design, sample size, comparison conditions, intervention duration, measurement approaches, and the extent to which findings could support causal or generalizable claims. This distinction is important because the absence of a formal risk-of-bias assessment does not imply that the included studies were considered methodologically equivalent. Rather, methodological limitations were incorporated into the interpretation of findings. For example, the quasi-experimental design and single-country setting of Ekizoğlu and Demir (2025) restrict generalizability, whereas the correlational design employed by Woo et al. (2025) does not permit causal conclusions concerning the effects of AI-generated text on writing quality.

Methodological and Contextual Limitations

Several methodological limitations were considered when interpreting the evidence;

First, the final evidence base consisted of only two studies, meaning that the conclusions drawn from the synthesis should be treated as preliminary rather than definitive.

Second, the studies were drawn from only Turkey and Hong Kong, leaving substantial geographical and educational contexts unrepresented. The findings therefore cannot be assumed to apply uniformly across secondary EFL/ESL systems.

Third, the two studies used fundamentally different research designs. Ekizoğlu and Demir (2025) employed a quasi-experimental comparison between AI-assisted and teacher-mediated feedback, whereas Woo et al. (2025) examined associations between AI-generated text use

and composition quality. This heterogeneity prevents direct quantitative comparison and limits the extent to which the findings can be interpreted as evidence of a common intervention effect.

Fourth, neither included study provided sufficient longitudinal evidence to establish whether observed improvements in writing performance, confidence, or composition quality persist beyond the immediate study or intervention period. Consequently, the review distinguishes between short-term product-level improvement and durable independent writing development.

Fifth, the use of a single reviewer for searching, screening, and verification introduces potential selection and interpretation bias.

Sixth, the reliance on a general academic search environment rather than comprehensive searches of Web of Science, Scopus, ERIC, and related specialist databases may have resulted in relevant studies being missed.

Finally, two potentially relevant Chinese studies could not be included because their authorship could not be independently verified. Their exclusion means that the final two-study evidence base should not be interpreted as necessarily representing the complete population of secondary-level studies published to date.

Ethical Considerations

The present study constituted a review of previously published empirical literature and did not involve the recruitment of participants, administration of interventions, collection of primary human-subject data, or processing of identifiable participant information. Accordingly, institutional ethical approval was not required for the review itself. Nevertheless, ethical considerations were incorporated into the interpretation of the literature because the population of interest comprises predominantly adolescent learners, many of whom are minors. Particular attention was therefore given to issues concerning teacher mediation, learner proficiency, responsible use of GenAI, and the potential risks of uncritical reliance on AI-generated feedback. The review also recognized that methodological quality and ethical safeguards in primary studies involving minors are important components of responsible GenAI research. For example, the reviewed evidence indicates that appropriate parental consent and institutional ethical approval were obtained in the Turkish secondary-school study. Such safeguards are particularly important when AI technologies are introduced into educational settings involving younger learners.

RESULTS

Overview of the Included Evidence

The search and source-verification process identified two studies that met the eligibility criteria and could be independently verified against their original publications: Ekizoğlu and Demir (2025) and Woo et al. (2025). The two studies were conducted in different educational contexts and employed substantially different methodological approaches. Ekizoğlu and Demir (2025) used a quasi-experimental design involving 60 Turkish secondary-school students and compared AI-generated feedback with traditional teacher feedback over an eight-week intervention. Woo et al. (2025), in contrast, examined 59 secondary-school students from seven schools in Hong Kong and investigated the relationship between AI-generated text use, student-authored text, revision, and composition quality. Because the studies differed in design, comparison conditions, outcome measures, and analytical approaches, their findings were not statistically pooled. Instead, the evidence was synthesized descriptively and thematically according to the three research questions guiding the review.

Research Question 1: What does the available secondary-school evidence indicate about the effects of AI-generated feedback and AI-generated text on writing accuracy, writing quality, and revision behavior?

The evidence indicates that AI-mediated writing support may be associated with improvements in selected dimensions of secondary-school learners' writing, but the pattern is not uniform across learners or forms of AI use. Ekizoğlu and Demir (2025) provided the clearest direct comparison of AI-generated and teacher-generated feedback. Their quasi-experimental study involved 60 Turkish secondary-school students divided between an experimental group receiving feedback from an AI writing assistant and a control group receiving traditional teacher feedback. Following the eight-week intervention, the AI-feedback group demonstrated significantly greater improvement in overall writing performance than the teacher-feedback group. The most notable differences were reported in grammar, vocabulary, and coherence. These findings provide evidence that AI-generated feedback can contribute to measurable improvements in several dimensions of writing performance under the conditions examined in the study. The finding, however, should not be interpreted as demonstrating that AI-generated feedback is universally superior to teacher feedback. The study was conducted with a relatively small sample in a single national context and over a limited intervention period. Furthermore, the comparison examined AI-generated feedback and teacher feedback as separate conditions rather than examining AI feedback as a supplement to teacher feedback. The result therefore demonstrates the potential effectiveness of AI-generated feedback within the specific instructional conditions investigated, rather than establishing a general

superiority of AI over teacher-mediated feedback. Woo et al. (2025) approached the issue from a different perspective. Rather than comparing AI-generated feedback with teacher feedback, the study examined how AI-generated text was incorporated into student compositions and how this related to writing quality. The findings indicated that the total number of words in a composition, regardless of whether the words were AI-generated or student-authored, was the strongest predictor of composition quality. This finding complicates a simple interpretation of AI-generated text as the direct cause of improved writing quality. It suggests that the relationship between AI assistance and writing performance cannot be reduced to the amount of AI-generated material incorporated into a composition.

The study further demonstrated that the relationship between AI-generated text and composition quality varied according to academic banding. Students from higher-banded schools tended to use AI-generated text more selectively and strategically, while lower-banded students demonstrated different patterns of dependence and revision. In particular, among lower-banded students, attempts to revise AI-generated text into their own words were sometimes associated with lower composition scores. This finding is important because it indicates that revising AI-generated material does not necessarily produce better writing outcomes for all learners.

The two studies provide evidence of potential benefits of AI-mediated writing support while simultaneously demonstrating substantial variation in outcomes. The Turkish study provides evidence of improvements in grammar, vocabulary, coherence, and overall writing performance following AI-generated feedback, whereas the Hong Kong study indicates that the relationship between AI-generated text, revision, and writing quality is more complex and may vary according to learners' academic banding. The available evidence therefore does not support a straightforward conclusion that greater use of AI-generated material necessarily produces better writing.

Research Question 2: How do secondary-school EFL/ESL learners engage with AI-generated feedback and AI-generated text, both behaviorally and affectively?

The two studies provide evidence of both positive affective responses and differentiated behavioral engagement with AI-generated writing support. Ekizoğlu and Demir (2025) reported a positive affective response among learners who received AI-generated feedback. Specifically, 87% of students in the experimental group reported increased confidence in their writing following the intervention. The authors associated this finding with increased writing self-efficacy. This result suggests that AI-generated feedback may contribute not only to changes in observable writing performance but also to learners' perceptions of their own writing ability. However, the finding should be interpreted cautiously because the study did not report a parallel self-efficacy measure for the teacher-feedback control group. It is therefore not possible to determine conclusively whether the increase in confidence resulted specifically from the AI component or from the broader experience of receiving structured writing feedback during the intervention. Woo et al. (2025) provides more differentiated evidence concerning learners' behavioral engagement with AI-generated text. The study showed that students did not interact with AI-generated material in a uniform manner. Higher-banded students tended to use AI-generated text more strategically, integrating it selectively while maintaining substantial student authorship. Their engagement therefore appeared to involve a more controlled and selective incorporation of AI-generated material.

The pattern among lower-banded students was different. Greater reliance on AI-generated text was associated with higher scores in some cases, but attempts to revise AI-generated text into students' own words were sometimes associated with lower composition scores. This finding is particularly important for understanding engagement because it suggests that increased learner involvement with AI-generated material does not automatically translate into stronger writing outcomes. The findings therefore indicate that learner engagement with AI-generated writing support is heterogeneous. Engagement may involve confidence-building and productive use of AI-generated feedback for some learners, while for others it may involve greater dependence on AI-generated material or difficulties in transforming AI-generated language into independently produced writing. From the perspective of the present review, these findings suggest that the educational value of AI-generated feedback cannot be determined solely by whether learners use the technology. The manner in which learners interact with AI-generated feedback or text appears to be an important consideration. The evidence also suggests that learner proficiency or academic banding may influence whether engagement with AI-generated material is productive.

Research Question 3: What methodological, geographical, educational, and contextual gaps characterize the emerging evidence on AI-mediated writing support among secondary-school EFL/ESL learners?

The evidence base identified in this review is notably limited in both size and geographical coverage. Only two studies satisfied the eligibility and source-verification criteria, and these were conducted in Turkey and Hong Kong. Consequently, the evidence does not represent the wider range of secondary-school EFL/ESL contexts in which AI-mediated writing support is being introduced.

The two studies also demonstrate considerable methodological heterogeneity. Ekizoğlu and Demir (2025) employed a quasi-experimental design with an AI-feedback group and a teacher-feedback comparison

group, whereas Woo et al. (2025) used a different approach centered on the relationship between AI-generated text use and composition quality. The studies consequently addressed related but not identical questions and used different outcome measures. This heterogeneity limits direct comparison between findings and prevents the calculation of a common intervention effect from the available evidence.

A further methodological gap concerns longitudinal evidence. The included studies do not establish whether observed improvements in writing performance or confidence persist over extended periods. This distinction is particularly important in relation to written corrective feedback because improvement in the immediate quality of a written product does not necessarily demonstrate durable language development or the independent internalization of linguistic knowledge. The evidence base is also limited in its examination of learner-level differences. Woo et al. (2025) provides evidence that academic banding may influence how students use and revise AI-generated material, with different patterns emerging among higher- and lower-banded students. However, the small number of available studies prevents a robust determination of how proficiency, motivation, prior AI experience, or other learner characteristics shape the effectiveness of AI-generated feedback.

There is also a significant gap concerning the distinction between AI-generated feedback and AI-generated text. Ekizoğlu and Demir (2025) directly investigated AI-generated feedback, whereas Woo et al. (2025) examined students' incorporation and revision of AI-generated text. Although both studies are relevant to AI-mediated writing support, they represent different forms of AI involvement in writing. Treating them as equivalent interventions would therefore obscure important differences in how AI enters the writing process.

The search process additionally identified two potentially relevant studies involving Chinese secondary-school learners: one concerning engagement with generative-AI feedback among middle-school students and another examining AI-generated and teacher-generated model texts among high-school students. These studies were not incorporated into the formal evidence base because their complete authorship information could not be independently verified. Their identification nevertheless indicates that the literature may be somewhat larger than the two studies formally included in this review and highlights the need for more comprehensive future searching and source verification. Overall, the methodological and contextual evidence indicates that secondary-school research on AI-mediated writing support remains at an early stage of development. The available studies are few, geographically concentrated, methodologically diverse, and limited in longitudinal evidence. These characteristics make it premature to draw broad conclusions about the effectiveness of AI-generated feedback for secondary-school EFL/ESL learners.

Summary of Findings across the Three Research Questions

Across the three research questions, the findings reveal a pattern of potential benefit combined with substantial conditionality. The first study indicates that AI-generated feedback can produce measurable improvements in grammar, vocabulary, coherence, and overall writing performance, while the second demonstrates that the relationship between AI-generated text, revision, and writing quality is more complex and varies across academic bands. The evidence concerning learner engagement similarly points to both positive and cautionary patterns. AI-generated feedback was associated with increased writing confidence among most learners in the Turkish study, while the Hong Kong study demonstrated that learners differed in how strategically they incorporated and revised AI-generated text.

Finally, the evidence base itself constitutes a major finding of the review. With only two verified studies from Turkey and Hong Kong, the literature remains too limited and heterogeneous to support broad generalizations concerning secondary-school EFL/ESL learners. The absence of substantial longitudinal evidence, the limited geographical coverage, differences in research design, and the distinction between AI-generated feedback and AI-generated text all indicate that further empirical research is required.

DISCUSSION

Interpreting the Evidence through Written Corrective Feedback Theory

The emerging evidence on AI-mediated written corrective feedback provides a modest but increasingly informative empirical basis for reconsidering the longstanding debate between Truscott (1996) and Ferris (1999) in a technologically transformed and developmentally distinct context. The significance of this literature lies not simply in whether generative artificial intelligence can identify and correct linguistic errors, but in whether AI-mediated correction produces the kinds of sustained linguistic development that have historically been at the center of the WCF debate. The distinction is important because the technological capacity to provide immediate and extensive corrective information does not, in itself, establish that learners have acquired the underlying linguistic knowledge necessary to transfer such corrections to subsequent writing. The contemporary evidence therefore permits the original debate to be reframed around a more differentiated question: under what conditions does AI-generated corrective feedback translate into durable improvements in second-language writing? At the broader level of WCF research, recent evidence provides a substantial qualification to a strongly

skeptical interpretation of corrective feedback. Brown et al. (2026), using a Bayesian meta-analytic approach across 52 primary studies with control groups, found robust evidence for moderate WCF effects that remained evident across short-, medium, and long-term measurement periods. Importantly, direct, indirect, and metalinguistic forms of WCF yielded broadly similar effect sizes across different research contexts, writing tasks, error types, and instructional characteristics. This evidence challenges a categorical interpretation of Truscott's (1996) position by demonstrating that corrective feedback can, under appropriate conditions, contribute to durable improvements in L2 writing accuracy. At the same time, the finding that WCF effects are moderate rather than transformative cautions against interpreting correction as a sufficient mechanism for comprehensive language development. The evidence therefore lends greater support to Ferris's (1999) rejection of an absolute dismissal of corrective feedback while simultaneously reinforcing the need to investigate the conditions and mechanisms through which feedback becomes educationally consequential. (Brown et al., 2026).

Evidence from automated WCF further strengthens this qualified interpretation. Rahimi et al. (2025), in a sequential explanatory mixed-methods study, compared EFL learners receiving electronic automated written corrective feedback with learners receiving non-electronic written corrective feedback. The electronic-feedback group outperformed the comparison group in overall writing performance, task achievement, and grammatical range and accuracy. However, no significant between-group differences were found in coherence and cohesion or lexical development. This pattern is theoretically important because it indicates that the effectiveness of automated feedback may be dimension-specific rather than uniformly distributed across writing proficiency. The findings consequently provide support for the proposition that AI-mediated corrective feedback can facilitate measurable improvement while simultaneously challenging any assumption that correction will produce equivalent gains across all components of writing. The qualitative evidence of behavioral, cognitive, and affective engagement with the automated feedback further suggests that the effects of feedback may depend not only on its provision but also on how learners interact with and respond to it. (Rahimi et al., 2025). This emphasis on engagement is particularly relevant when the Ferris–Truscott debate is reconsidered in relation to generative AI. If correction is to contribute to durable development, learners must do more than encounter the corrected form; they must engage with the feedback in ways that support interpretation, evaluation, revision, and subsequent application. Rahimi et al. (2025) provide evidence that such engagement can occur in an automated feedback environment, but their findings also demonstrate that engagement does not necessarily translate into uniform improvement across all dimensions of writing. This distinction is consistent with the central concern underlying Truscott's (1996) skepticism: the pedagogical problem is not simply whether errors can be corrected, but whether correction contributes to underlying language development. Contemporary AI systems may make corrective information more accessible and immediate, but the educational significance of that information remains dependent upon the learner's engagement with it.

The timing of feedback provides another important qualification to any straightforwardly optimistic account of AI-mediated correction. Cheng and Zhang (2025), comparing synchronous and asynchronous computer-mediated WCF among Chinese tertiary EFL learners, found that both forms significantly improved linguistic accuracy, while synchronous feedback was more effective than asynchronous feedback. However, neither treatment produced corresponding improvements in syntactic complexity or fluency. These findings suggest that the temporal immediacy of feedback may influence particular aspects of linguistic performance without necessarily generating broad-based improvements in writing proficiency. This is particularly relevant to generative AI, whose defining affordances include rapid and continuous interaction with learners. Although immediate feedback may increase the likelihood that a linguistic discrepancy is addressed while the writing task remains cognitively accessible, immediacy should not be equated with durable acquisition. The evidence instead suggests that the timing of corrective intervention is one factor that may condition its effectiveness. (Cheng & Zhang, 2025).

The role of individual learner characteristics provides a further complication. Ni and Xu (2025), drawing on Control-Value Theory, found that English proficiency had only a limited influence on EFL learners' emotional responses to WCF, whereas learning motivation played a much more substantial role. Highly motivated learners experienced both positive and negative emotions towards corrective feedback more intensely than their less motivated counterparts, while instrumental motivation demonstrated a stronger association with emotional responses than integrative motivation. These findings are significant because they broaden the interpretation of feedback effectiveness beyond observable changes in writing accuracy. Feedback is experienced by learners, and those experiences may shape how they attend to, evaluate, accept, resist, or act upon corrective information. In an AI-mediated environment, where learners may interact with feedback with greater independence from teachers, motivational and affective characteristics may become particularly consequential. Thus, even technically accurate feedback may have different pedagogical consequences depending on the learner's disposition towards learning and feedback itself. (Ni & Xu, 2025). The emergence of generative AI nevertheless introduces a technological dimension that distinguishes contemporary automated feedback from

earlier forms of computer-mediated correction. Wu et al. (2025) found that GPT-4, particularly when supported by task-specific prompting, significantly outperformed commercial automated writing tools in the consistency and accuracy of error detection and correction. Their empirical evidence further suggested potential improvements in students' writing quality and confidence and possible reductions in teacher workload. These findings provide important support for the proposition that generative AI can overcome some of the limitations associated with earlier automated WCF systems. However, the reported concern regarding occasional overcorrection is equally important. If an AI system modifies forms that are acceptable or pedagogically appropriate, learners may be exposed to corrections that appear authoritative but are not necessarily warranted. The problem therefore shifts from whether AI can provide feedback to whether learners can critically evaluate the validity and relevance of feedback provided by AI. (Wu et al., 2025).

This issue has particular significance when the evidence is interpreted through the Ferris–Truscott debate. The capacity of generative AI to provide feedback consistently and at scale appears to address one of the practical limitations of teacher-mediated WCF: teachers cannot necessarily provide immediate, detailed correction on every linguistic feature of every learner's draft. From a Ferris-oriented perspective, the scalability and consistency of AI-generated feedback could strengthen opportunities for learners to receive and respond to corrective information. However, from a Truscott-oriented perspective, increased correction does not necessarily solve the fundamental problem of whether learners develop durable linguistic competence. Indeed, if learners accept AI-generated revisions without understanding why they are appropriate, the apparent efficiency of automated correction could potentially conceal rather than resolve the underlying learning problem. The technological superiority of a feedback system in detecting errors should therefore be analytically distinguished from its pedagogical effectiveness in promoting independent language development. The emerging secondary-school evidence makes this distinction even more consequential. Ekizoğlu and Demir (2025) reported that AI-assisted writing feedback produced greater gains than teacher feedback alone in grammar, vocabulary, and coherence among Turkish secondary-school students, suggesting that AI-mediated feedback can produce measurable improvements in several dimensions of adolescent EFL writing. However, Woo et al. (2025) found a considerably more differentiated pattern among Hong Kong secondary-school students. Their findings indicated that the relationship between AI-generated text and composition quality varied according to students' academic banding, and that revising AI-generated text in their own words was sometimes associated with lower composition scores among lower-banded students. The juxtaposition of these findings are particularly revealing. It suggests that AI-mediated writing support cannot be evaluated independently of learner characteristics and the manner in which learners engage with AI-generated language.

The findings of Woo et al. (2025) are especially important in relation to the durability question at the center of the WCF debate. If lower-proficiency learners reproduce, revise, or otherwise manipulate AI-generated language without possessing sufficient linguistic resources to evaluate the output, the resulting improvement in a particular written product may not necessarily represent corresponding development in their independent writing competence. This possibility resonates with Truscott's (1996) concern that correction without genuine learner development may have limited long-term value. However, it would be equally inappropriate to interpret the finding as evidence that AI-generated feedback is inherently ineffective. Rather, the result indicates that the relationship between corrective intervention and writing development may be moderated by learner proficiency and the nature of learner engagement with the feedback. This interpretation is more consistent with the contemporary WCF evidence, which increasingly points towards conditional rather than universal effects. The evidence therefore complicates any attempt to position the contemporary literature wholly within either the Ferris or Truscott camp. Brown et al. (2026) provide robust evidence that WCF can have durable effects, thereby challenging a categorical rejection of corrective feedback. Rahimi et al. (2025) demonstrate that automated feedback can improve overall writing performance, task achievement, and grammatical accuracy while producing less consistent effects on coherence, cohesion, and lexicon. Cheng and Zhang (2025) show that the timing of feedback can influence accuracy outcomes without necessarily improving syntactic complexity or fluency. Ni and Xu (2025) demonstrate that learner motivation shapes emotional responses to feedback, highlighting the importance of individual differences. Wu et al. (2025) demonstrate that generative AI can substantially enhance the technical performance of automated correction while also introducing risks such as overcorrection. Finally, the emerging secondary-school evidence indicates that these effects may not be distributed uniformly across adolescent learners.

The theoretical implication is therefore not that the Ferris–Truscott debate has been settled by generative AI, but that the technological transformation of feedback exposes the limitations of framing the debate as a binary question of whether correction "works". The contemporary evidence suggests that the more analytically productive question concerns what type of feedback works, for whom, under what conditions, through which mechanisms, and with what consequences for subsequent independent writing. Generative AI potentially increases the frequency, immediacy, consistency, and scalability of corrective intervention, but these affordances do not eliminate the cognitive,

affective, and developmental conditions necessary for feedback to become learning. For the present review, this distinction is particularly important because the available secondary-school evidence remains limited relative to the rapidly expanding higher-education literature. The two directly relevant secondary-school studies do not provide sufficient grounds for a general conclusion about the effectiveness of AI-generated feedback among adolescent EFL learners. Rather, they reveal a pattern of conditionality that warrants systematic investigation. The positive findings reported by Ekizoğlu and Demir (2025) indicate that AI-mediated feedback may support measurable gains in several dimensions of writing, whereas Woo et al. (2025) demonstrate that the outcomes of AI engagement may vary according to learner academic banding and the manner in which AI-generated language is incorporated into student writing. These findings should therefore be treated as complementary evidence of heterogeneity rather than as mutually exclusive conclusions.

Accordingly, the present review adopts a position that moves beyond both technological optimism and technological skepticism. AI-generated WCF is conceptualized here as a potentially powerful but inherently mediated pedagogical intervention. Its effectiveness is likely to depend upon the interaction between the characteristics of the feedback system, the timing and nature of feedback delivery, learner proficiency and motivation, the extent and quality of learner engagement, the writing dimension being targeted, and the instructional context in which AI is deployed. This positioning preserves the central insight of Ferris (1999)—that corrective feedback can contribute to L2 writing development—while retaining the critical caution associated with Truscott (1996): correction should not be equated automatically with durable learning. The contribution of the present review is therefore to examine whether, and under what conditions, generative AI can transform corrective feedback from an act of error identification into a process that supports sustained and transferable development in secondary-school EFL writing.

Interpreting the Evidence through the Noticing Hypothesis

The Noticing Hypothesis fares similarly unevenly against this evidence. Ekizoğlu and Demir (2025) explained their experimental group's gains partly through the immediacy of AI generated feedback prompting greater noticing of errors than delayed teacher feedback typically allows, a reading consistent with Schmidt's (1990) original claim that conscious attention to input is necessary for it to become intake. Yet Woo et al. (2025) found that for lower proficiency students, active engagement with AI generated text, arguably the clearest behavioral indicator of noticing available in their data, was sometimes associated with weaker rather than stronger outcomes. One plausible explanation, though one this review's evidence base cannot confirm directly, is that noticing alone may be insufficient for lower proficiency learners who lack the linguistic resources to act productively on what they notice, suggesting that the Noticing Hypothesis may require additional scaffolding conditions to hold at lower levels of secondary learner proficiency than it does for the more linguistically developed populations on which much of the hypothesis's original supporting research was conducted.

The Uneven Nature of Benefit

A central pattern across both studies, and arguably the most important finding this review is positioned to offer, is that AI generated feedback's benefits at the secondary level are not uniform. This unevenness is consistent with the broader feedback literature beyond second language writing specifically; Hattie and Timperley's (2007) influential synthesis of feedback research found that feedback's effectiveness depends heavily on the specific conditions under which it is delivered and received, rather than operating as a uniformly beneficial intervention regardless of context, a finding that anticipates exactly the kind of moderation by learner proficiency and school context observed in the present review's charted secondary level studies. Ekizoğlu and Demir's (2025) sixty student sample showed a clear average advantage for the AI feedback condition, but the study did not examine whether this average effect masked variation across students of differing proficiency within the experimental group itself. Woo et al.'s (2025) finding that benefit was moderated by school academic banding suggests strongly that such variation likely exists and likely matters. Read together, these findings caution against treating “AI generated feedback works for secondary learners” as a settled or uniform claim, and instead point toward a more precise and more useful question for future research, namely which secondary learners, under which conditions, benefit from which forms of AI generated feedback.

Ethical Considerations for Secondary Level AI Use

The finding that benefit is uneven across proficiency levels carries direct implications for how AI generated feedback should be introduced to secondary classrooms responsibly. As discussed earlier in this review, secondary learners are minors in most jurisdictions, and Ekizoğlu and Demir's (2025) explicit attention to parental consent and institutional ethics approval reflects an appropriate baseline standard of care. The evidence charted here suggests an additional, more specific ethical consideration, namely that lower proficiency secondary learners, who may be the students with the greatest need for effective feedback, also appear to be the students for whom AI generated text engagement carries the most uncertain or potentially adverse relationship with writing

outcomes. This suggests that AI generated feedback may be most safely and effectively introduced to secondary classrooms under teacher mediation rather than as a tool placed directly and without scaffolding in the hands of the lowest proficiency learners, a position consistent with the broader ethical preference, discussed in the methodology of this review, for mediated rather than standalone generative AI use with adolescent learners.

Limitations of this Review

This review carries several limitations that should temper confidence in its conclusions. Most significantly, the charted evidence base consists of only two sources, drawn from two national contexts. While this scarcity is itself an honest and informative finding about the state of secondary level research, it also means that the patterns identified here, including the proficiency moderated nature of benefit, rest on a thin empirical foundation and should be treated as hypotheses for future research rather than established conclusions. The review's search process relied on a general academic web search tool rather than formal subscription databases, and was conducted by a single reviewer without independent dual screening, both of which depart from systematic review convention, though both are more readily accommodated within scoping review methodology. Two additional candidate studies identified during the search process could not be included due to unverifiable authorship, meaning the true secondary level evidence base may be marginally larger than what this review formally charted. Future research extending this review should prioritize formal database searches across Scopus, Web of Science, and ERIC, independent dual screening with a reported inter-rater reliability statistic, and direct verification of the two studies excluded here on authorship grounds alone.

Conclusion

This review set out to address a specific and verifiable gap in the literature on AI generated feedback in EFL and ESL writing instruction, namely the near total absence of secondary level evidence within a research base otherwise concentrated in higher education, a gap explicitly acknowledged by Lo et al. (2024) and implicitly reflected in the broader scope of Crosthwaite and Sun's (2025) review. The two studies meeting this review's eligibility and verification criteria, Ekizoğlu and Demir (2025) and Woo et al. (2025), offer an early and still limited picture, but one with a clear and pedagogically significant pattern, namely that AI generated feedback's benefits for secondary learners appear genuine but uneven, shaped by learner proficiency and school context in ways that the higher education literature has not needed to grapple with to the same degree. For EFL and ESL educators working with adolescent learners, the most defensible current position is one of cautious, mediated integration rather than either wholesale adoption or wholesale avoidance of AI generated feedback, paired with close attention to how individual students, particularly those with lower proficiency, actually engage with the tool rather than reliance on average effects alone. For researchers, the priority is clear, namely a substantially larger and more methodologically diverse body of secondary level evidence, gathered through formal systematic search and verified dual screening, capable of testing whether the patterns identified in this early evidence base hold across the wide range of educational systems and contexts in which secondary EFL and ESL learners encounter generative AI today.

REFERENCES

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

Bogdanov, S. (2026). Reconceptualizing Noticing in Second Language Acquisition: Task Relevance, Effort, and Evidential Limits. English Studies at NBU, 12(1), 133-149.

Brown, D., Liu, Q., & Norouzian, R. (2026). Effectiveness of written corrective feedback in developing L2 accuracy: A Bayesian meta-analysis. Language Teaching Research, 30(3), 1357-1389.

Cheng, X., & Zhang, L. J. (2025). Investigating synchronous and asynchronous written corrective feedback in a computer-assisted environment: EFL learners' linguistic performance and perspectives. Computer Assisted Language Learning, 38(7), 1687-1716.

Crosthwaite, P., & Sun, S. (2025). Generative AI and L2 written feedback studies: A scoping review. RELC Journal. Advance online publication. https://doi.org/10.1177/00336882251386530

Ekizoğlu, M., & Demir, A. N. (2025). The role of AI assisted writing feedback in developing secondary students writing skills. Discover Education, 4, Article 454. https://doi.org/10.1007/s44217-025-00919-3

ELkhettm, S., & Habiba, M. (2026). The Study of Noticing and Input Hypotheses in Advancing EFL Learners' Performance. Transcultural Journal of Humanities and Social Sciences, 7(1), 24-69.

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). Journal of Second Language Writing, 8(1), 1–11.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Jensen, L. X., Bearman, M., Boud, D., & Konradsen, F. (2026). Feedback encounters in doctoral supervision: The role of generative AI chatbots. Assessment & Evaluation in Higher Education, 51(5), 849-862.

Juwita, R., Lestari, P., Lestari, L., Lestari, V. A., & Mutawally, A. N. (2025). The Role of Gemini Ai in Facilitating Noticing and Retention of Language Features: A Descriptive Quantitative Analysis. Innovex, 1(1), 69-74.

Lindsay, E. D., Zhang, M., Johri, A., & Bjerva, J. (2025, April). The responsible development of automated student feedback with generative AI. In 2025 IEEE Global Engineering Education Conference (EDUCON) (pp. 1-10). IEEE.

Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S. Y. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. Smart Learning Environments, 11, Article 50. https://doi.org/10.1186/s40561-024-00342-5

Ni, F., & Xu, W. (2025). How do English proficiency and learning motivation shape EFL students' emotions toward written corrective feedback. System, 131, 103681.

Otaki, B. (2023). Feedback in the era of generative AI.

Rahimi, M., Fathi, J., & Zou, D. (2025). Exploring the impact of automated written corrective feedback on the academic writing skills of EFL learners: An activity theory perspective. Education and Information Technologies, 30(3), 2691-2735.

Saini, A. K., Tzirides, A. O., Cope, B., Kalantzis, M., & Searsmith, D. (2025). Generative Ai and the Next Generation of Feedback Systems: A Hybrid Approach for Higher Education. In Proceedings of the 19th International Conference of the Learning Sciences-ICLS 2025, pp. 2120-2124. International Society of the Learning Sciences.

Sajadi, S., Huerta, M. A. R. K., Ryan, O., & Drinkwater, K. (2024). Harnessing generative AI to enhance feedback quality in peer evaluations within project-based learning contexts. International Journal of Engineering Education, 40(5), 998-1012.

Schmidt, R. (1990). The role of consciousness in second language learning. Applied Linguistics, 11(2), 129–158.

Shahini, A. (2025). The noticing hypothesis in language learning. In Language Learning Theories: A Student's Guide (pp. 103-120). Cham: Springer Nature Switzerland.

Szcześniak, K., Tsikandilakis, M., Toranzos, R. L., & Dömischová, I. (2026). Picked up from seeded input: the Noticing Hypothesis and exposure to binomials. Lingua, 341, 104219.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. Language Learning, 46(2), 327–369.

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, Article 1249991. https://doi.org/10.3389/fpsyg.2023.1249991

Woo, D. J., Susanto, H., Yeung, C. H., & Guo, K. (2025). Approaching the limits to EFL writing enhancement with AI-generated text and diverse learners [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.00367.

Wu, J., Li, J., Ge, Z., Xu, M., Lin, L., & Zhang, R. (2025). Effectiveness of generative AI in automated written corrective feedback with prompting. Journal of Educational Computing Research, 63(6), 1493-1527.

Zhang, K., Li, S., & Huang, Y. (2026). Generative AI-driven feedback in higher education: a scoping review. Assessment & Evaluation in Higher Education, 1-18.