



Review Article

Generative AI-Mediated Feedback and L2 Grammatical-Lexical Development: A Scoping Review of Longitudinal Research Gaps

Le Thi Thuan An¹ , Nguyen Xuan Hong^{2,*}

¹Faculty of Foreign Languages, Ba Ria - Vung Tau University, Ho Chi Minh City, Viet Nam

²Faculty of Foreign Languages, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Viet Nam

Abstract

The rapid adoption of generative artificial intelligence (AI) tools such as ChatGPT in second language (L2) writing instruction has produced a fast-growing body of research on AI-generated written corrective feedback. Two recent scoping reviews have mapped this literature broadly, focusing on feedback quality, learner uptake and pedagogical integration. However, neither review specifically examined how this literature treats grammatical and lexical development as distinct, measurable constructs, nor how it addresses the moderating role of learner proficiency over time. This scoping review synthesizes 30 empirical and methodological sources, identified through a targeted, PRISMA-ScR-informed search of Google Scholar, Scopus-indexed journal content and ERIC, together with citation networks, to map the state of longitudinal research on generative-AI-mediated feedback and grammatical-lexical development in L2/EFL writing. The review finds that most intervention studies remain limited to a single semester or shorter, that grammatical accuracy and lexical development are rarely measured as separate, co-tracked outcomes and that only a small number of very recent studies (2025-2026) have applied longitudinal growth-modelling techniques, none of which disentangle grammar from vocabulary or systematically test proficiency level as a moderator of developmental trajectories. No such study appears to have been conducted in the Vietnamese EFL context, based on the sources identified in this review. The review proposes a research agenda for longitudinal, multi-proficiency-level, construct-differentiated research on generative-AI-mediated feedback and outlines implications for instructional design and future primary research.

Keywords

Generative AI, Written Corrective Feedback, L2 Writing, Grammatical Accuracy, Lexical Diversity

1. Introduction

Written corrective feedback (WCF) is widely regarded as one of the central mechanisms supporting second language (L2) writing development and decades of research and meta-analysis

have examined its effectiveness, moderators and boundary conditions [1-4]. The public release of ChatGPT in November 2022 introduced a qualitatively new source of WCF, namely a con-

*Correspondence: Nguyen Xuan Hong (nguyenxuanhong@iuh.edu.vn)

Received: 9 September 2026; Accepted: 18 September 2026; Published: 30 September 2026



Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

versational, generative artificial intelligence (AI) system capable of producing immediate, individualised feedback across multiple dimensions of writing at essentially unlimited scale [5]. Within roughly three years, research on generative-AI-mediated feedback in L2 writing has expanded from a handful of exploratory case studies into a literature large enough to itself require synthesis.

Two recent scoping reviews have begun this synthesis work. One review [6] mapped the early trajectory of generative AI and L2 written feedback research, documenting its rapid growth and the preliminary, largely exploratory nature of the evidence base. More recently, a considerably larger-scale scoping review of 185 empirical studies published between November 2022 and March 2026 was conducted [7], organized around feedback type, learner uptake, comparative effectiveness and pedagogical integration. That review identified four influential patterns. First, comparative AI-human feedback studies dominate the field. Second, content-level and higher-order feedback remain comparatively underexamined. Third, learner uptake is measured as a primary outcome in only a small minority of studies. Fourth, generative AI feedback performs comparably to teacher feedback on surface-level accuracy but less consistently on content and argumentation, with learner perceptions of AI feedback often exceeding its measured effect on performance.

These two reviews provide an invaluable map of the field's overall shape, but neither was designed to answer a narrower and, for L2 development research specifically, more consequential question: how has this literature treated the linguistic constructs of grammatical accuracy and lexical development and to what extent has it employed the longitudinal, repeated-measures designs needed to track how these constructs change over time and across learners at different proficiency levels? Grammatical accuracy and lexical richness are not interchangeable indicators of a single underlying "writing quality" factor. L2 acquisition theory treats them as at least partially dissociable systems that may develop along different timelines and respond differently to the same pedagogical input [8, 9]. A synthesis focused specifically on this construct-level and temporal dimension of the literature is therefore a necessary complement to the broader scoping reviews already available.

This review accordingly asks four questions. (1) How has generative-AI-mediated feedback research operationalised and measured grammatical accuracy and lexical development in L2 writing? (2) To what extent have these two constructs been examined jointly, within the same study and the same learners? (3) To what extent has this research employed longitudinal or growth-modelling designs capable of tracking developmental trajectories over time? (4) To what extent has learner proficiency level been examined as a moderator of these trajectories and in what geographic and institutional contexts? Answering these questions maps precisely the gap that a longitudinal, multi-proficiency-level study of generative-AI-mediated feedback and grammatical-lexical development, of the kind proposed in the companion research proposal to

this article, would need to be addressed.

2. Method

2.1. Design

This review follows the scoping review framework [10], with the procedural refinements proposed in [11], and is reported in line with the PRISMA extension for scoping reviews (PRISMA-ScR [12]). A scoping design was chosen because the aim is to map the extent, distribution and characteristics of a specific slice of a fast-moving literature, its treatment of grammatical and lexical development and its use of longitudinal designs, rather than to pool effect sizes or appraise the certainty of a body of intervention evidence, for which a systematic review or meta-analysis would be the more appropriate tool. Given that two broader scoping reviews of the general generative-AI-feedback literature already exist [6, 7], this review is deliberately narrower in scope: rather than re-mapping the entire field, it targets the intersection of three specific features such as grammatical accuracy as an outcome, lexical development as an outcome and longitudinal or repeated-measures design that the broader reviews did not use as an organizing lens.

2.2. Search Strategy

A structured literature search was conducted using Google Scholar, Scopus-indexed journal content and ERIC (Education Resources Information Center), together with backward citation searching of the reference lists of the two identified scoping reviews [6, 7] and of the highest-relevance primary studies identified. Search terms combined three concept clusters using Boolean operators. The first cluster covered technology terms: "generative AI", "ChatGPT", "GPT-4" and "large language model (LLM)". The second cluster covered feedback and writing terms: "written corrective feedback", "feedback", "L2 writing", "EFL writing" (English as a Foreign Language) and "ESL writing" (English as a Second Language). The third cluster covered construct or design terms: "grammatical accuracy", "lexical diversity", "lexical development", "longitudinal", "growth curve", "developmental trajectory" and "proficiency level". The search covered publications from November 2022 (the public release of ChatGPT) through August 2026 and was not restricted by publication type, allowing inclusion of both peer-reviewed journal articles and rigorously reported conference proceedings. The three concept clusters were combined using the syntax native to each platform (for example, in Scopus and ERIC: TITLE-ABS-KEY (("generative AI" OR "ChatGPT" OR "GPT-4" OR "large language model") AND ("written corrective feedback" OR feedback OR "L2 writing" OR "EFL writing" OR "ESL writing") AND ("grammatical accuracy" OR "lexical diversity" OR "lexical development" OR longitudinal OR "growth curve" OR "developmental trajectory" OR "proficiency level"))); Google Scholar searches used equivalent free-text combinations of the

same three clusters, screened manually because the platform does not support field-restricted Boolean strings. Screening proceeded in two stages: titles and abstracts were first screened against the eligibility criteria in Section 2.3, followed by full-text assessment of all records retained at the first stage; both stages were conducted by the authors, with disagreements resolved through discussion.

2.3. Eligibility Criteria

Sources were included if they met three criteria. First, they reported original empirical data or a methodological/theoretical synthesis directly relevant to generative-AI-mediated feedback in L2/EFL writing. Second, they addressed grammatical accuracy, lexical development or both as an outcome or explicitly employed a longitudinal or repeated-measures design in this domain. Third, they were available in English with a verifiable author, venue and, where applicable, a DOI (Digital Object Identifier). Sources were excluded if they concerned feedback on speaking rather than writing, addressed automated writing evaluation (AWE) tools that predate generative AI (e.g., Grammarly, Criterion) without a generative-AI comparison condition or could not be bibliographically verified. Foundational theoretical and methodological works predating the generative-AI era (e.g., core WCF theory, feedback-literacy theory, lexical-diversity measurement) were included where they provide the conceptual or methodological scaffolding for interpreting the generative-AI literature.

2.4. Study Selection and Charting

The search and citation-tracing process identified a working corpus of 30 sources meeting the eligibility criteria: 21 empirical studies of generative-AI-mediated feedback in L2 writing, 2 existing scoping reviews of the broader field, 1 meta-analysis of written corrective feedback effectiveness and 6 foundational theoretical or methodological references. Each empirical source was charted along six dimensions. The first was publication year. The second was learner population and geographic/institutional context. The third was the generative-AI tool used. The fourth was the outcome construct(s) measured (grammatical accuracy, lexical development, both or neither as a primary focus). The fifth was study duration and number of measurement waves. The sixth was whether proficiency level was examined as a moderating variable. This charting exercise, summarized in [Table 1](#), directly structures the thematic synthesis presented in Section 3. Consistent with scoping review methodology, no formal risk-of-bias or effect-size pooling was undertaken. Where effect sizes or descriptive statistics are reported below, they are drawn from individual studies and presented as illustrative rather than pooled estimates.

2.5. Synthesis Approach

A narrative synthesis was conducted, organized thematically around the review's four guiding questions. Given the

modest and heterogeneous size of the corpus such as spanning experimental, quasi-experimental, case-study and mixed-methods design, a narrative rather than a quantitative synthesis was judged most appropriate, consistent with the approach taken in the two broader scoping reviews on which this review builds [\[6, 7\]](#).

3. Results

3.1. Overview of the Literature Landscape

Consistent with the two existing broad-scope reviews, the corpus charted here confirms an accelerating publication trajectory: only a handful of studies appeared in 2023, growing to a substantial majority of the identified corpus published in 2024 and 2025, with a further cluster already appearing in the first half of 2026 [\[7\]](#). ChatGPT remains the overwhelmingly dominant tool studied, frequently without a specified model version, which limits model-level comparison [\[7\]](#) given that GPT-3.5, GPT-4, GPT-4o, Claude and Gemini differ substantially in capability. Most studies in the present corpus are quasi-experimental or case-study designs conducted within a single institution and a single semester, a pattern that recurs throughout the thematic synthesis below.

The charting of the broader field in [\[7\]](#) is informative for contextualizing the present, narrower synthesis. Across that 185-study corpus, general or mixed feedback was the most frequently studied feedback type (40.5%), followed by comparative AI-human designs (27.6%). Corrective, form-focused feedback, the category most relevant to grammatical accuracy, accounted for a smaller 18.4% of the field and content-level or holistic feedback for a still smaller 13.5%. Learner uptake, that is, whether and how learners actually act on the feedback they receive, was charted as a primary outcome in only a small minority of studies. This distribution matters for the present review because it suggests that even the broader literature on which grammatical- and lexical-development research depends is itself skewed toward describing feedback delivery and comparative quality rather than tracking what learners do with that feedback linguistically over time.

How have outcomes been operationalized? A closely related pattern concerns measurement choice. Across the studies charted for this review, three broad approaches to outcome operationalization recur. The first, by far the most common, is a holistic or composite writing-quality score, often a rubric covering task achievement, organization, language use and mechanics collapsed into a single number, used in comparative studies such as [\[13\]](#) and in sequencing studies such as [\[14\]](#). The second is a discrete linguistic index applied to a specific construct, exemplified by the use of the Measure of Textual Lexical Diversity (MTLD) and the Moving-Average Type-Token Ratio (MATTR) for lexical diversity [\[15, 16\]](#) or manual error-taxonomy coding for grammatical accuracy [\[16\]](#). The third, increasingly visible in the very newest studies, is a latent

variable estimated through growth-curve modelling from repeated observed indicators [17-19]. These three approaches are not evenly distributed across the corpus: composite scores dominate cross-sectional and short intervention studies, discrete linguistic indices appear almost exclusively in the small number of corpus-linguistics-informed studies such as [16] and latent growth modelling appears only in the three longitudinal studies discussed in Section 3.5. No study in the corpus was found to combine the second and third approaches. That is, no study feeds discrete, construct-specific linguistic indices such as MTLT or an error-free-clause ratio into a latent growth model to estimate their developmental trajectories separately over time.

3.2. Generative AI Feedback and Grammatical Accuracy

Grammatical accuracy is the outcome most consistently and successfully targeted by generative-AI feedback research, reflecting both the relative ease of operationalizing surface-level correctness and the well-documented strength of large language models at local error detection. A mixed-method multiple case study of four L2 writers over five weeks [5] found that ChatGPT-provided automated written corrective feedback (AWCF) yielded correct-revision rates for grammatical errors exceeding those reported in earlier research using Grammarly as the feedback source, a pattern attributed to ChatGPT's comparative strength in grammatical error correction. Similarly, ChatGPT-4 was reported to achieve high accuracy in identifying grammatical and lexical issues [20], though with lower specificity for higher-order concerns such as task response and coherence. In a rigorously designed quasi-experimental comparison [13], no statistically significant difference was found between AI-generated and teacher-generated feedback on argumentative writing performance overall (Cohen's $d = 0.10$), even though both conditions produced significant within-group gains, an important null finding given the study's comparatively strong design [7].

These generative-AI-specific findings sit within a much larger prior literature on written corrective feedback in general, which has consistently found feedback to be moderately effective for improving grammatical accuracy. A meta-analysis of 21 primary studies [4] estimated an overall effect size of $g = 0.68$ for the effect of WCF on L2 grammatical accuracy, while also identifying learner proficiency as one of the strongest moderators of feedback efficacy, with more advanced learners benefiting more consistently than beginners. Notably, this proficiency-moderation finding predates the generative-AI literature entirely. It derives from decades of teacher- and peer-feedback research and as the synthesis in Section 3.6 shows does not appear, based on the sources identified in this review, to have been systematically re-examined for generative-AI feedback specifically, let alone within a design capable of tracking how the moderating effect of proficiency unfolds over time.

A further complication for interpreting grammatical-accuracy findings across this literature is that accuracy gains are not always applied consistently once feedback is delivered. One study [21] found that ChatGPT-suggested grammatical revisions were not always retained across successive drafts and that new errors occasionally emerged during more complex sentence restructuring, indicating that a single-point accuracy measurement can understate the volatility of grammatical development under generative-AI feedback conditions. Relatedly, a large-scale comparison of ChatGPT and teacher feedback [22], drawing on 1,200 feedback records from 60 EFL secondary students, found a pronounced asymmetry in how learners acted on feedback depending on its linguistic level, surface-level versus meaning-level. Large grammatical feedback was acted on at a substantially higher rate (approximately 86% for ChatGPT and 92% for teachers) than meaning-level feedback concerning organization or ideas (approximately 32% and 29%, respectively). Although this study did not measure lexical diversity as a separate construct, its surface-versus-meaning distinction is conceptually adjacent to the grammar-versus-lexis distinction this review is concerned with and suggests that uptake itself, not only feedback delivery, may differ systematically by linguistic level, a possibility that a longitudinal design tracking both constructs could test directly.

3.3. Generative AI Feedback and Lexical Development

Lexical development has received comparatively less direct attention than grammatical accuracy and is more often assessed as a secondary or exploratory outcome than as a primary focus of study design. Where it has been measured, the Measure of Textual Lexical Diversity (MTLD [15]) has emerged as the preferred index in this domain, owing to its demonstrated stability across varying text lengths relative to older indices such as the type-token ratio. A comparison of pure generative-AI feedback with hybrid AI-plus-teacher feedback among 60 Chinese EFL students over a 12-week period [23] reported that AI feedback alone was particularly effective at accelerating lexical refinement and surface-level accuracy gains, whereas hybrid feedback produced comparatively stronger gains in deeper, discourse-level revision. This finding suggests generative AI feedback may have a comparative advantage specifically in the lexical and grammatical domain relative to higher-order writing concerns, a pattern broadly consistent with the field-wide observation [7] that generative AI feedback performs more reliably on surface-level than content-level dimensions of writing.

Beyond this, however, lexical development remains a comparatively thin strand of the literature. Most studies that mention vocabulary do so as part of a composite "writing quality" or holistic rubric score rather than through a dedicated, validated lexical-diversity metric, which limits the precision with which lexical gains attributable to generative-AI feedback can be distinguished from gains in other writing dimensions. For

example, one study [24] found that ChatGPT tended to generate a substantially greater volume of feedback than teachers across content, organization and language combined, but its coding scheme did not isolate lexical suggestions from grammatical or organizational ones, so the specific contribution of AI feedback to vocabulary growth cannot be recovered from that data. A similar pattern is evident in an otherwise fine-grained analysis of ChatGPT-4's feedback specificity [20], where grammatical and lexical dimensions were reported together as a single "language use" category achieving higher specificity than task-response or coherence feedback, again precluding a separate estimate of lexical-feedback quality alone. This recurring conflation of grammar and lexis under broader "language" or "accuracy" categories is arguably the single most consistent measurement gap identified across the corpus and is precisely what the study discussed next [16] was designed to overcome.

3.4. Studies Jointly Examining Grammatical and Lexical Outcomes

Only one study identified in this review's corpus measured grammatical accuracy and lexical diversity as two explicit, separately operationalized outcomes within the same generative-AI feedback study. One study [16] examined 91 in-class ESL texts alongside their revised versions following teacher-moderated ChatGPT feedback, coding accuracy using an established error taxonomy and computing lexical diversity using both MTLTD and MATTR. It found that lexical diversity increased in the ChatGPT-revised texts relative to the in-class versions, with MTLTD scores rising from a mean of approximately 78 to approximately 86, alongside accuracy gains, concluding that ChatGPT-mediated revision can support integrated, simultaneous gains across both linguistic dimensions in a way that pre-generative-AI automated writing evaluation tools were not well equipped to provide.

This study is an important proof of concept. It demonstrates empirically that grammatical accuracy and lexical diversity can be measured jointly and can respond differently or with different magnitudes to the same generative-AI feedback intervention. However, its design is a single-session, cross-sectional comparison between an in-class draft and one AI-mediated revision, rather than a repeated-measures design tracking how these two constructs co-develop over multiple writing tasks and multiple weeks or months. It therefore establishes the empirical and methodological feasibility of jointly tracking grammar and lexis, without answering the longitudinal question of how their developmental trajectories unfold, converge or diverge over an extended period of instruction, precisely the gap that a longitudinal design would address.

3.5. The Emergence of Longitudinal and Growth-Modelling Designs

Until very recently, the generative-AI-feedback literature

was almost entirely cross-sectional or limited to short interventions of a few weeks. It has been noted that the field as a whole shows a marked absence of longitudinal outcome research [7], and the short duration of a five-week design was explicitly identified as its principal limitation [5], calling for future longitudinal investigations of the long-term effects of generative-AI feedback on learner behaviour and outcomes.

This review's search identified an important and very recent shift: a small cluster of studies published in late 2025 and 2026 have begun applying latent growth curve modelling (LGC/LGCA), a structural-equation-modelling technique that estimates both the average developmental trajectory of an outcome over time and the extent to which individuals or groups deviate from it, to generative-AI feedback research. What appears to be the first such study [17] longitudinally compared generative-AI feedback, exemplar-based feedback and teacher feedback over a 12-week period among Chinese undergraduates, using latent growth curve analysis to model trajectories of L2 writing development and examining gender as a moderating variable. Its findings indicated that AI-generated and exemplar-based feedback were associated with more sustained developmental trajectories than teacher feedback alone. In a related study [18], the same longitudinal growth-modelling approach was applied to a different but related outcome, student feedback literacy, finding that generative-AI, peer and exemplar feedback modalities were associated with different trajectories of feedback-literacy development among 162 Chinese undergraduates. This growth-modelling approach was extended further still [19], tracking the trajectory of L2 writing enjoyment (an affective rather than a linguistic outcome) across three time points in a single semester among 328 participants, again using latent growth curve modelling.

These three studies collectively establish that longitudinal, growth-modelling designs are methodologically feasible and are beginning to be adopted in this research area. However, none of them addresses the specific gap this review is concerned with. The study in [17] models "L2 writing development" as a global outcome rather than disentangling grammatical accuracy from lexical development as separate growth processes. Its chosen moderator is gender rather than proficiency level. Its design also spans 12 weeks, equivalent to one semester, rather than a full academic year. Moreover, that study, like the other two growth-modelling studies identified, was conducted with Chinese undergraduates, with no equivalent study identified in this review's search in a Vietnamese or, more broadly, Southeast Asian EFL context. Table 1 summarises how each longitudinal or joint-construct study charted in this review relates to the specific combination of features such as construct differentiation, proficiency-level moderation, extended duration and Vietnamese context that remains unaddressed.

3.6. Proficiency Level as a Moderator

As noted in Section 3.2, proficiency level is well established in the pre-generative-AI WCF literature as one of the

strongest moderators of feedback effectiveness [4], and Skill Acquisition Theory [8, 9] provides a strong theoretical basis for expecting learners at different stages of declarative-to-automatised knowledge to respond differently to the same corrective input. Within the generative-AI-specific literature, however, this variable has been examined only sporadically and rarely as a primary focus. One study [13] tested for an interaction between feedback type (AI vs. teacher) and proficiency level and found none, although learners at an intermediate-low level showed the largest within-group gains under both feedback conditions, a finding that hints at, without confirming, a possible curvilinear relationship between proficiency and responsiveness to feedback. Beyond this single study, no source identified in this review examined proficiency level as a moderator within a longitudinal or growth-modelling design; on the basis of the literature located here, it remains unclear whether learners at different starting proficiency levels exhibit different developmental slopes, different rates of acceleration or plateauing or different optimal feedback dosages over an extended period of generative-AI-mediated instruction.

Individual-difference research adjacent to proficiency offers indirect support for expecting such heterogeneity. One study [5] found that higher-proficiency and more technologically competent learners refined their ChatGPT prompts iteratively and evaluated suggested feedback critically before revising, whereas lower-proficiency learners engaged more superficially and revised primarily at the surface level. A study of graduate ESL students' screencasted revision behaviour [25] similarly found that participants relied on ChatGPT chiefly for lower-order concerns such as paraphrasing and local correction even after being trained in more sophisticated prompting strategies, despite reporting high satisfaction with the tool. This pattern is consistent with the feedback-literacy findings synthesized in [7], in which evaluative judgement and metacognitive awareness, rather than mere access to feedback, predicted the depth and productiveness of learner uptake. These findings suggest that proficiency level is likely to shape not only how much grammatical or lexical gain a learner realizes from generative-AI feedback, but also how selectively and critically that learner engages with the feedback in the first

place. A longitudinal design is uniquely positioned to disentangle these two possibilities. A multi-wave, multi-cohort study can examine whether proficiency-related differences in outcome trajectories are better explained by differences in linguistic readiness to benefit from correction, by differences in feedback-engagement behaviour or by an interaction between the two that changes as learners themselves develop over the course of a year.

3.7. Southeast Asian and Vietnamese Contexts

The geographic distribution of the charted corpus is heavily weighted toward East Asian (particularly Chinese), Middle Eastern and Western European institutional contexts. Within Southeast Asia, one study [14] offers a directly relevant data point: a 15-week study of 14 Vietnamese undergraduates examining the sequencing of AI-generated and teacher-generated feedback, which found that AI-first feedback sequences were associated with stronger local, sentence-level revisions, while teacher-first sequences better supported higher-order content revision. This study demonstrates both that generative-AI feedback research is beginning to reach the Vietnamese EFL context and that existing Vietnamese-context work remains limited to a single semester, a small sample and holistic revision-quality outcomes rather than validated, separately tracked measures of grammatical accuracy and lexical diversity. No study located through this review's search has been found to combine a Vietnamese or comparable Southeast Asian EFL population with a multi-wave longitudinal design and construct-differentiated (grammar and lexis) outcome measurement.

3.8. Synthesis: Mapping the Research Gaps

Table 1 consolidates the charting exercise described in Section 2.4 for the studies most central to this review's four guiding questions, mapping each against the four defining features of the gap this review has identified: differentiated grammar/lexis measurement, a longitudinal or growth-modelling design, proficiency level as an examined moderator and a Vietnamese/Southeast Asian research context.

Table 1. Mapping of core studies against the review's four gap-defining features.

Study	Grammar + lexis differentiated?	Longitudinal / growth model?	Proficiency as moderator?	Vietnamese/SE Asian context?
Yan & Zhang (2024)	Grammar only	No (5 weeks, single wave outcome)	No (descriptive only)	No
Zhang, Aubrey, Huang & Chiu (2025)	Both mentioned, not separately modelled	No (12 weeks, pre/post)	No	No
Alnemrat et al. (2025)	No (holistic score)	No	Yes (tested, ns interaction)	No

Study	Grammar + lexis differentiated?	Longitudinal / growth model?	Proficiency as moderator?	Vietnamese/SE Asian context?
Pretorius & Thewissen (2026)	Yes (both measured)	No (single-session revision)	No	No
Zhu, Wang & Qin (2025)	No (global writing development)	Yes (LGCA, 12 weeks)	No (gender examined instead)	No
Zhu, Qin, Wang & Liu (2026)	No (feedback literacy, not linguistic)	Yes (LGCA)	No	No
Shi et al. (2026)	No (writing enjoyment, affective)	Yes (LGCM, 3 waves/1 semester)	No	No
Tran (2025)	No (holistic revision quality)	No (15 weeks, cross-sectional comparison)	No	Yes (Vietnam)
Proposed study	Yes (grammar and lexis as separate growth processes)	Yes (1 academic year, 4-6 waves)	Yes (3 proficiency cohorts)	Yes (Vietnam)

Reading across this table rather than down any single row is the central finding of this review: every one of the four features has been demonstrated as individually feasible somewhere in the recent literature, including construct differentiation [16], longitudinal growth modelling [17], proficiency-level testing [13] and a Vietnamese research context [14], but no identified study combines more than two of these features at once and none combines all four. The gap this review identifies is therefore not a gap in any single method or measure, each of which now has a demonstrated precedent, but a gap in their integration.

4. Discussion

4.1. Why Grammar and Lexis Need to be Disentangled

The tendency, documented throughout Section 3, to treat "writing quality" or "writing development" as a single composite outcome is understandable from a practical standpoint. Holistic rubrics are easier to apply and often more directly aligned with high-stakes assessment. However, this practice obscures a theoretically important possibility: that grammatical and lexical systems may respond to generative-AI feedback along different timelines. Skill Acquisition Theory [8, 9] predicts that grammatical rule learning proceeds through declarative, procedural and automatised stages that depend on repeated, feedback-informed practice, while models of lexical growth (as reviewed in the context of MTLTD's development [15]) treat vocabulary breadth and depth as accumulating somewhat more continuously through exposure and use. If these two systems do indeed follow different growth curves under generative-AI feedback, then studies that collapse them into a single score risk masking exactly the kind of differential, theoretically diagnostic finding that would most advance the

field. For instance, grammatical accuracy might plateau after an initial period of rapid AI-assisted gain while lexical diversity continues to increase steadily or vice versa. The finding that lexical diversity increased alongside accuracy gains in a single-session comparison [16] is a first hint that the two constructs can be jointly tracked and may show distinguishable patterns, but a single time point cannot reveal whether their trajectories diverge over an extended period.

This distinction is not merely methodological pedantry. From a pedagogical standpoint, a finding that grammatical accuracy and lexical diversity develop in lockstep under generative-AI feedback would justify continuing to treat "language use" as a single instructional target, consistent with how most rubrics in the charted corpus currently operate. A finding that the two constructs diverge would instead argue for differentiated instructional emphasis and potentially for different feedback protocols (direct versus metalinguistic, in the terms of [26]) targeted at each construct. This could happen, for instance, if repeated exposure to direct, error-specific AI correction accelerates grammatical gains while contributing comparatively little to lexical range or if the reverse pattern holds. Because no longitudinal study in the corpus has been designed to detect such divergence, this remains an open, empirically answerable and pedagogically consequential question rather than a settled matter of theoretical preference.

4.2. Why Proficiency-Level Moderation Matters for Longitudinal Design

The absence of proficiency-level moderation from the longitudinal generative-AI literature is arguably the more consequential gap of the two identified in this review. The meta-analytic finding that proficiency is among the strongest moderators of WCF effectiveness in the pre-generative-AI literature [4], combined with a single, cross-sectional finding of

curvilinear-looking (though statistically non-significant) proficiency-related gain patterns [13], together suggest that a learner's starting proficiency level plausibly shapes not just how much they benefit from generative-AI feedback, but also how that benefit unfolds over time. For example, lower-proficiency learners might show a slower initial uptake followed by accelerating gains as declarative knowledge proceduralises, while higher-proficiency learners might reach a ceiling more quickly. None of the growth-modelling studies identified in Section 3.5 examined this possibility. Instead, they selected gender [17] or no moderator at all [19] as their between-person variable of interest. A multi-cohort longitudinal design that follows learners at multiple, clearly defined starting proficiency levels in parallel is therefore positioned to answer a question that is theoretically well-motivated but empirically untested.

4.3. Implications for a Longitudinal, Multi-Proficiency Research Agenda

Taken together, the synthesis in Section 3 supports three concrete implications for future primary research. Each implication corresponds directly to a design feature of the longitudinal study proposed in the accompanying research proposal. First, future studies should measure grammatical accuracy and lexical development as separate, validated outcomes rather than defaulting to a composite writing-quality score. This follows the precedent set methodologically in [16]. Second, future studies should adopt repeated-measures or growth-modelling designs extending across at least one full academic year, building on the recently demonstrated feasibility of latent growth curve approaches [17-19] while extending their typical 12-week/one-semester duration. Third, future studies should treat proficiency level as a primary, *a priori* moderator rather than as an incidental control variable. This is ideally done through a multi-cohort design that follows learners at several distinct proficiency bands in parallel. Future studies should also extend this line of enquiry into under-represented contexts such as Vietnam. As Section 3.7 shows, the existing evidence base there remains limited to short, single-cohort, holistically scored studies [14].

A fourth, more design-specific implication concerns the choice of feedback protocol itself. Although hybrid AI-teacher feedback models fall outside this review's core focus on construct and design gaps, the broader literature charted in [7] found that such hybrid models consistently outperformed either AI-only or teacher-only feedback. These hybrid models accounted for only 12.4% of that corpus. The sequencing of AI and teacher input also shaped which dimensions of writing improved most. This has a direct bearing on how a future longitudinal grammar-lexis study should specify its intervention. Rather than treating generative-AI feedback as a simple substitute for teacher feedback, researchers should document precisely which feedback types (direct, indirect or metalinguistic, following the typology in [26]) are delegated to the AI system

and which, if any, remain with the instructor. This specification is itself likely to influence the shape of the grammatical and lexical growth trajectories observed.

5. Limitations of This Review

This review has several limitations. First, its search, while structured and informed by PRISM A-ScR reporting principles, was not an exhaustive, database-native systematic search of the kind conducted in [7] across a full Scopus export. Some relevant studies may therefore not have been captured, particularly recent preprints, regional-language publications or studies indexed outside the databases and citation networks consulted. Readers seeking a comprehensive map of the entire generative-AI-feedback literature should consult [7] and [6] directly. The present review is deliberately narrower and is intended to complement rather than replace those broader syntheses. Second, because this is a scoping rather than a systematic review, no formal risk-of-bias appraisal or effect-size pooling was conducted. The illustrative statistics reported in Section 3 (e.g., specific effect sizes, MTLTD values) should therefore be read as characteristics of individual studies rather than as pooled, generalisable estimates. Third, the field itself is evolving extremely rapidly. Three of the most directly relevant studies identified in this review were published within the twelve months preceding its completion. As a result, any synthesis of this literature carries an unusually short shelf life and should be periodically revisited as new longitudinal studies emerge.

6. Conclusion and Implications for Future Longitudinal Research

This scoping review set out to map a specific, theoretically consequential slice of the rapidly growing literature on generative-AI-mediated feedback in L2 writing: its treatment of grammatical accuracy and lexical development as distinguishable outcomes and its use of longitudinal designs capable of tracking how these outcomes develop over time and across learners at different proficiency levels. Building on two recent broad-scope scoping reviews of this literature [6, 7], the present synthesis of 30 sources shows that generative-AI feedback research has, until very recently, been dominated by short, single-semester, cross-sectional designs measuring a composite "writing quality" outcome. A small but growing cluster of 2025-2026 studies has begun to apply longitudinal growth-modelling techniques to this domain. A separate, methodologically important 2026 study has also demonstrated that grammatical accuracy and lexical diversity can be jointly and validly measured within a single generative-AI feedback study. However, no identified study combines all four features together: a longitudinal, growth-modelling design, separately tracked grammatical and lexical outcomes, proficiency level as a primary moderator and a Vietnamese or comparable

Southeast Asian EFL research context. A longitudinal, multi-cohort study designed explicitly to integrate these four features, of the kind detailed in the accompanying research proposal, would therefore address a gap that is well motivated by existing theory [8, 9, 27] and precisely delimited by the empirical literature synthesized here. Such a study would also extend both the construct-level precision and the geographic reach of a fast-moving and increasingly consequential field of applied linguistics research.

More broadly, this review's core contribution is methodological rather than purely descriptive. It charts the corpus against four specific, jointly necessary design features: construct differentiation, longitudinal growth modelling, a priori proficiency moderation and an under-represented research

context, rather than against the field's overall size or thematic breadth. This approach shows that the apparent maturity of the generative-AI-feedback literature, now numbering in the hundreds of studies [7], can coexist with a narrow, well-defined and still entirely open gap at the intersection of construct, design and context. Researchers planning primary studies in this area and the Vietnamese doctoral/PhD-candidate track or doctoral applicants scoping a dissertation topic within it, would therefore be well advised to look past the field's raw publication volume. Instead, they should ask which specific combinations of construct, design and context remain untested. This review has attempted to answer that question precisely for the case of generative-AI-mediated feedback and L2 grammatical-lexical development.

Abbreviations

AI	Artificial Intelligence
AWCF	Automated Written Corrective Feedback
AWE	Automated Writing Evaluation
https	//doi.org/Digital Object Identifier
EFL	English as a Foreign Language
ERIC	Education Resources Information Center
ESL	English as a Second Language
L2	Second Language
LGCA	Latent Growth Curve Analysis
LGCM	Latent Growth Curve Modelling
LLM	Large Language Model
MATTR	Moving-Average Type-Token Ratio
MTLD	Measure of Textual Lexical Diversity
NCS	Nghiên C?u Sinh (Vietnamese doctoral/PhD-candidate Track)
PRISMA-ScR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
WCF	Written Corrective Feedback

Author Contributions

Le Thi Thuan An: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology

Nguyen Xuan Hong: Resources, Supervision, Validation, Writing – original draft, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Bitchener, J., Ferris, D. R. Written corrective feedback in second language acquisition and writing. New York, NY: Routledge; 2012.
- [2] Bitchener, J., Storch, N. Written corrective feedback for L2 development. Bristol, UK: Multilingual Matters; 2016.
<https://doi.org/10.21832/9781783095056>
- [3] Ferris, D. R. Response to student writing: Implications for second language students. Mahwah, NJ: Lawrence Erlbaum; 2003.
- [4] Kang, E., Han, Z. The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. The Modern Language Journal. 2015, 99(1), 1-18.
<https://doi.org/10.1111/modl.12189>
- [5] Yan, D., Zhang, S. L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. Humanities and Social Sciences Communications. 2024, 11, 1086.
<https://doi.org/10.1057/s41599-024-03543-y>
- [6] Crosthwaite, P., Sun, S. Generative AI and L2 written feedback studies: A scoping review. RELC Journal. 2025, 57(1), 207-219.
<https://doi.org/10.1177/00336882251386530>

- [7] Craven, L., Fredrick, D. R. LLM-generated feedback in L2 writing: A scoping review. *Education Sciences*. 2026, 16(8), 1196. <https://doi.org/10.3390/educsci16081196>
- [8] DeKeyser, R. M. Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*. 1997, 19(2), 195-221.
- [9] DeKeyser, R. M., Suzuki, Y. Skill acquisition theory. In: Van-Patten B, Keating GD, Wulff S, editors. *Theories in second language acquisition: An introduction*. 4th ed. New York, NY: Routledge; 2025, p. 157-182.
- [10] Arksey, H., O'Malley, L. Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*. 2005, 8(1), 19-32. <https://doi.org/10.1080/1364557032000119616>
- [11] Levac, D., Colquhoun, H., O'Brien, K. K. Scoping studies: Advancing the methodology. *Implementation Science*. 2010, 5(1), 69. <https://doi.org/10.1186/1748-5908-5-69>
- [12] Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C.,... Straus, S. E. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*. 2018, 169(7), 467-473. <https://doi.org/10.7326/M18-0850>
- [13] Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., AlSharefeen, R. AI vs. teacher feedback on EFL argumentative writing: A quantitative study. *Frontiers in Education*. 2025, 10, 1614673. <https://doi.org/10.3389/feduc.2025.1614673>
- [14] Tran, T. T. T. Enhancing EFL writing revision practices: The impact of AI- and teacher-generated feedback and their sequences. *Education Sciences*. 2025, 15(2), 232. <https://doi.org/10.3390/educsci15020232>
- [15] McCarthy, P. M., Jarvis, S. MTLD, vocd-D and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*. 2010, 42(2), 381-392. <https://doi.org/10.3758/BRM.42.2.381>
- [16] Pretorius, M., Thewissen, J. Mediating L2 writing revision: The impact of teacher-moderated ChatGPT feedback on L2 accuracy and lexical diversity. *Journal of Second Language Writing*. 2026, 71, 101286. <https://doi.org/10.1016/j.jslw.2026.101286>
- [17] Zhu, R., Wang, H., Qin, X. Longitudinal comparison of AI, exemplar and teacher feedback for sustainable L2 writing development: A latent growth curve analysis. *Innovation in Language Learning and Teaching*. 2025. <https://doi.org/10.1080/17501229.2025.2586142>
- [18] Zhu, R., Qin, X., Wang, H., Liu, X. Longitudinal effects of feedback from peer, AI and exemplars on student feedback literacy development: A latent growth curve analysis. *Assessment and Evaluation in Higher Education*. 2026. <https://doi.org/10.1080/02602938.2026.2631543>
- [19] Shi, L., Zuo, Y., Li, Z. Longitudinal analysis of L2 writing enjoyment trajectory in AI-mediated writing: Examining the roles of AI feedback literacy and learner-AI interactivity. *European Journal of Education*. 2026, 61(3), e70726. <https://doi.org/10.1111/ejed.70726>
- [20] Saricaoglu, A., Bilki, Z. The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity and relevance. *Annual Review of Applied Linguistics*. 2025, 45, 253-273. <https://doi.org/10.1017/S0267190525100160>
- [21] ElEbyary, K., Shabara, R. ChatGPT-generated corrective feedback: Does it do what it says on the tin? *Teaching English with Technology*. 2024, 24(3), 68-89. <https://doi.org/10.56297/vaca6841/BFFO7057/MYEH4562>
- [22] Yu, H., Xie, Q. Generative AI vs. teachers: Feedback quality, feedback uptake and revision. *Language Teaching Research Quarterly*. 2025, 47, 113-137. <https://doi.org/10.32038/ltrq.2025.47.07>
- [23] Zhang, Z., Aubrey, S., Huang, X., Chiu, T. K. F. The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. *Innovation in Language Learning and Teaching*. 2025. <https://doi.org/10.1080/17501229.2025.2503890>
- [24] Guo, K., Wang, D. To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*. 2024, 29(7), 8435-8463. <https://doi.org/10.1007/s10639-023-12146-0>
- [25] Koltovskaia, S., Rahmati, P., Saeli, H. Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive and affective engagement. *Journal of Second Language Writing*. 2024, 65, 101130. <https://doi.org/10.1016/j.jslw.2024.101130>
- [26] Ellis, R. A typology of written corrective feedback types. *ELT Journal*. 2009, 63(2), 97-107. <https://doi.org/10.1093/elt/ccn023>
- [27] Vygotsky, L. S. *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press; 1978.