TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language

Pedagogically Structured AI Feedback and Grammatical Accuracy in EFL Speaking: A 14-Week Longitudinal Study

*****On the Internet*****

August 2026 – Volume 30, Number 2
https://doi.org/10.55593/ej.30118int1

Lilit Sargsyan
Yerevan State University, Armenia
<4lilitvs@gmail.com>

Abstract

Grammatical accuracy in spontaneous second language (L2) speech remains difficult to develop, particularly in contexts with limited opportunities for sustained oral practice. This study investigates whether structured, memory-enabled artificial intelligence (AI) speech coaching supports the development of grammatical accuracy among Armenian EFL learners.

A 14-week longitudinal mixed-methods design was employed with 64 undergraduate students, supplemented by a non-randomized comparison group (n = 13). Learners engaged in weekly AI-supported speaking tasks, and grammatical accuracy was measured as errors per 100 words across three time points.

Repeated-measures ANOVA revealed a significant reduction in grammatical errors over time (ηp^2 = .45), with the strongest improvements observed in rule-governed categories such as tense–aspect, subject–verb agreement, and word order, and more limited progress in articles and prepositions. Moderate-proficiency learners demonstrated the most consistent gains, while the comparison group showed no significant improvement.

The findings suggest that AI-mediated feedback can support grammatical development when embedded within structured pedagogical frameworks, while highlighting the importance of instructional guidance to prevent overreliance on automated corrections. This study contributes longitudinal evidence on AI-supported grammatical development in spontaneous speech and provides a fine-grained analysis of category-specific improvement.

Keywords: AI speech coaching; Grammatical accuracy in spontaneous speech; Armenian EFL learners; Longitudinal intervention; Corrective feedback

AI Feedback in EFL Speaking Development

Grammatical accuracy in spontaneous spoken English remains a persistent challenge for EFL learners, particularly in contexts where opportunities for sustained oral practice are limited. Memory-enabled AI tools, unlike static CALL systems, are responsive to longitudinal performance patterns, providing a level of continuity and individualization less commonly achievable in traditional classroom settings (Jaganov et al., 2025). This is particularly relevant for learners whose first language differs substantially from English in its grammatical structure.

However, research on AI-mediated second language development remains fragmented, with most studies focusing on fluency and confidence rather than structural accuracy in spontaneous speech. This study addresses this gap by investigating the extent to which persistent L1-driven errors can be reduced through sustained AI-supported speaking practice.

Spoken-Written Asymmetry in EFL Performance

Although differences between grammatical accuracy in writing and in speaking are well documented in second-language acquisition research (e.g., DeKeyser & Suzuki, 2025; Ellis, 2017), their implications for classroom-based speaking development remain underexplored. This issue is particularly evident in contexts such as Armenian higher education, where recent reductions in instructional hours appear to have limited in-class L2 exposure. As many universities still prioritize written forms of formative and summative assessment, students often rely on self-directed learning to develop their written accuracy, resulting in greater grammatical stability due to available time for monitoring. Conversely, as speaking unfolds in real time, learners are forced to conceptualize ideas, retrieve vocabulary, and assemble syntax under significant time pressure. This cognitive load often leads to inconsistencies in tense marking, word order and auxiliary use.

Preliminary classroom observations in such contexts suggest a recurring pattern, in which students who excel in written grammar often produce markedly less accurate spoken English. This asymmetry reflects learners' reliance on declarative knowledge of grammatical rules while struggling to deploy them procedurally in oral speech. These observations align with the distinction central to theoretical models such as Skill Acquisition Theory, according to which form-focused instruction without opportunities for proceduralization tends to result in inert knowledge (DeKeyser & Suzuki, 2025). Armenian learners often demonstrate extensive reliance on monitoring in writing but limited monitoring in real-time speech (Krashen, 1982).

These cognitive and curricular factors converge to create a compelling rationale for intervention. Specifically, interventions that combine high-frequency oral practice with targeted feedback are particularly needed where institutional assessment has historically emphasized written accuracy. However, existing research has not sufficiently examined how such asymmetries can be addressed through sustained, feedback-driven speaking practice, particularly in technology-supported environments.

Cross-linguistic differences play a crucial role in shaping patterns of grammatical error in second language production. Armenian, an independent branch of the Indo-European language family, differs significantly from English in its reliance on case marking rather than function words and in its comparatively flexible word order (Dum-Tragut, 2009). Cross-linguistic transfer theory (Jarvis, 2011; Odlin, 1989) predicts such error patterns when learners lack robust L2 proceduralization. Given these persistent cross-linguistic challenges, recent developments in AI offer new possibilities for targeted pedagogical intervention.

Memory-Enabled AI as Emerging Support for Spoken Grammar

Recent generative AI systems reference learners' previous outputs across sessions (OpenAI, 2026), enabling continuity in feedback rather than isolated corrections. This longitudinal memory can allow the system to detect persistent error patterns and adjust feedback accordingly. Throughout this article, we use "AI speech coach" to refer to a conversational system configured to provide accuracy-focused feedback, and "AI-mediated feedback" to describe the corrective information provided. Unlike traditional classroom instruction, which rarely affords sustained learner-specific feedback, AI coaches can deliver multiple correction cycles and track whether specific error types persist across weeks. Such systems also support distributed practice and repeated retrieval, which are known to enhance long-term retention (Cepeda et al., 2006).

In this study, pedagogical structuring refers to the deliberate design of prompts, feedback cycles, and constraints aimed at directing learners' attention to grammatical form during communicative practice.

Theoretical Framework

Skill Acquisition Theory (SAT) provides a widely used framework for understanding how learners develop grammatical accuracy in second language production. Within this framework, AI-mediated feedback may be conceptualized as supporting different stages of skill development, including explicit explanation during the declarative phase and repeated corrective feedback during proceduralization (DeKeyser & Suzuki, 2025). Limitations emerge when learners lack an underlying conceptual category, as with definiteness for article-less L1 speakers. Nevertheless, the integration of AI in weekly speaking tasks is conceptually aligned with SAT's emphasis on meaningful, contextualized practice.

In addition to Skill Acquisition Theory, Interactionist and Focus-on-Form frameworks provide complementary perspectives on how feedback supports language development. Within this tradition, Focus on Form emphasizes the integration of attention to linguistic features within communicative activity (Long, 1991), while interactionist research highlights the role of feedback moves such as recasts, prompts, and metalinguistic cues in facilitating noticing and uptake (Lyster & Ranta, 1997). AI-mediated systems appear capable of delivering hybrid forms of such feedback, including recasts, metalinguistic explanation, and prompts for self-correction. These frameworks jointly support the study's rationale. However, empirical evidence remains limited regarding whether such AI-mediated feedback leads to measurable improvements in spontaneous spoken grammatical accuracy over time, which the present study seeks to investigate.

Recent studies have increasingly explored the role of AI in supporting L2 speaking development. For example, Mingyan et al. (2025) report improvements in fluency and confidence following AI-mediated speaking practice in their recent study. This trend is also reflected in broader reviews of AI-supported informal language learning environments (Liu & Zhao, 2026). However, relatively few studies focus explicitly on grammatical accuracy in spontaneous speech. Additionally, studies reporting grammatical improvement often rely on written production or universal accuracy measures, without mapping the occurrence of category-specific grammatical change, which makes it difficult to identify how and why grammatical improvements occur. As a result, the understanding of how AI feedback affects specific grammatical categories in spontaneous speech remains limited.

Research Gap and Research Questions

Despite growing interest in AI-assisted language learning, research on grammatical accuracy in AI-mediated speaking remains limited. Existing studies continue to prioritize fluency and confidence, with comparatively less attention to structural accuracy in spontaneous speech. Furthermore, limited evidence exists on how AI feedback affects specific grammatical categories and how learner proficiency mediates these effects. The present study addresses these gaps through a longitudinal analysis of grammatical development in AI-supported speaking.

This study was guided by the following research questions:

- ***To what extent does sustained interaction with a memory-enabled AI speech coach improve the grammatical accuracy of EFL learners' spontaneous spoken English over time?***
- ***Do different categories of L1-influenced grammatical errors (e.g., tense–aspect, subject–verb agreement, word order, articles, and prepositions) show differential rates of improvement during AI-mediated speaking practice?***
- ***Does initial proficiency level moderate the effects of memory-enabled AI feedback on the development of spoken grammatical accuracy?***
- ***How does structured AI-mediated coaching compare with unstructured, self-directed AI use in improving grammatical accuracy in spoken English?***

This study makes three contributions. First, it provides longitudinal evidence of grammatical development in spontaneous spoken English under AI-mediated feedback. Second, it offers a fine-grained analysis of category-specific error reduction. Third, it examines how structured AI use interacts with learner proficiency and contrasts with unstructured use.

Method

Participants and the Integration of AI Tools

The study involved 64 first-year undergraduate EFL students enrolled in an Armenian university from the Faculties of Mathematics and Mechanics and of Economics and Management. All participants had prior EFL instruction but varied in oral proficiency. Based on a diagnostic speaking assessment, learners were grouped into low (n = 18), moderate (n = 29), and high (n = 17) proficiency levels. This classification enabled differential analysis across proficiency levels.

Participants interacted with a generative AI system (ChatGPT, OpenAI; GPT-4 with memory enabled and GPT-4 Turbo variants). A standardized prompt instructed the AI to evaluate grammatical accuracy, identify recurring errors, track changes across sessions, and gradually increase task difficulty. The system was configured to provide corrective feedback, including reformulations and metalinguistic explanations (see Appendix B).

Students completed a weekly speaking task by recording a short oral response and submitting it to the AI speech coach. After each submission, the AI analyzed the transcript, identified grammatical errors, and provided three forms of feedback: brief metalinguistic explanations, localized corrective reformulations, and prompts encouraging self-correction. Students were instructed to review the feedback carefully, compare the corrected segments with their original utterances, and ask follow-up questions when clarification was needed. They were also encouraged to revisit recurring feedback points before completing subsequent speaking tasks and to incorporate the

corrections into additional practice. With memory enabled, the AI retained summaries of recurring grammatical error patterns across sessions and generated reminders targeting patterns that persisted over time.

Learner reflections and submitted feedback logs indicated that many participants actively engaged with the feedback, although the depth of engagement varied across individuals.

Learners interacted with the AI outside class. Although the duration of AI interaction was not strictly controlled, students were instructed to engage in AI-supported speaking practice for approximately 30 minutes per week (typically divided into two sessions), ensuring a minimum level of exposure across participants.

Participants were informed about AI data handling practices and advised not to include personally identifying information. All participants provided informed consent in alignment with institutional ethical guidelines.

As the researcher also served as the course instructor, an independent research assistant was involved in transcript verification and error coding to mitigate potential bias. As students were encouraged to use AI-supported practice as part of the course design, they were also informed that their level of engagement and submission of feedback logs did not affect their grades.

The intervention is replicable at the pedagogical level, as its core components (e.g., accuracy-focused prompts, feedback cycles, and attention to recurring errors) can be implemented with similar AI systems.

An additional group of 13 second-year undergraduates served as a comparison group. These students had similar educational backgrounds but did not receive structured AI-supported feedback. They participated in comparable weekly speaking tasks.

Research Design

A quasi-experimental longitudinal mixed-methods design was employed, incorporating a non-randomized comparison group for contextual interpretation rather than causal inference. The research design, following Dörnyei (2007), combined quantitative error tracking with qualitative analysis of feedback logs. The study spanned 14 weeks and included three measurement points: pre-test (Week 1), mid-test (Week 7), and post-test (Week 14).

Qualitative data included AI feedback logs and learner reflections. Learners were provided with brief reflection prompts focusing on recurring errors, perceived improvement, and their interaction with AI feedback. Weekly speaking tasks ensured sustained exposure to AI-mediated correction.

The design of the study was aligned with the research questions, enabling analysis of overall grammatical accuracy development, category-specific error patterns, differences across proficiency levels, and comparisons between structured and unstructured AI use.

AI Memory Configuration

With memory enabled, the AI was configured to store representations of recurrent errors across sessions (OpenAI, 2026) while anonymizing individual utterances (OpenAI, 2026). During the study period (September 2024 - January 2025), the system transitioned from GPT-4 to GPT-4 Turbo. While memory functionality remained stable, subtle changes in feedback phrasing or error detection sensitivity may have occurred. This model instability is acknowledged as a study limitation.

Data Collection Procedures

Methodological challenges in AI-mediated speaking research have been widely noted, particularly regarding ASR accuracy (e.g., LoCamato & Munro, 2023). To address this, the current study employed manual verification: instructors reviewed all ASR transcripts, corrected misrecognitions, and confirmed genuine learner errors, reducing transcription-related distortion common in ASR-based SLA research. These limitations highlight the need for careful validation of spoken data, which informs the methodological design of the present study.

The data collection procedure included students submitting their weekly speaking responses via the Moodle platform. This allowed the instructor to monitor student work and adjust prompts or scaffolding as needed. Speaking tasks included impromptu discussions, short argumentative responses, and topic-based monologues designed to elicit spontaneous language production.

On a monthly basis, *students' speech samples were recorded* using personal smartphones or the Moodle Mobile App and saved as MP3 files. OpenAI's Whisper ASR system was used to automatically transcribe the recordings, which were then subjected to rigorous manual verification.

The transcript verification procedure had two stages:

- All transcripts were first verified by the instructor, who listened to each audio file and manually corrected inaccuracies. All incomprehensible segments were identified and eliminated.
- To evaluate reliability, 25% of randomly chosen transcripts were independently verified by a qualified research assistant. Strong agreement was indicated by Cohen's kappa (κ = .89, 95% CI [.85,.93]).

This two-stage procedure ensured high transcription accuracy and reduced false negatives in error counts, addressing a key methodological concern in ASR-based research.

Using a predefined coding framework, errors were classified into seven grammatical categories: articles, pronouns, prepositions, tense–aspect, subject–verb agreement, word order, and clause construction errors. For each utterance, the coders assigned an error category and provided a detailed description. The coders also recorded:

- Whether the error was "episodic" or "persistent" ($\geq$ 3 occurrences over weeks);
- Whether the error had already been detected by the AI's memory.

When assessing the advantages of AI memory, persistent errors were the main focus.

Measures and Focus on Accuracy

Measures of complexity, accuracy, and fluency (CAF) were initially derived. The present study focuses on grammatical accuracy, operationalized as errors per 100 words. This normalized measure is commonly used in L2 accuracy research because it controls for variation in production length and enables comparisons across learners and time points (Mizumoto, 2025; Polio & Shea, 2014). Although complexity and fluency metrics were also collected, the present analysis focuses solely on grammatical accuracy. This focused approach allows clearer analysis of how memory-enabled AI affects specific structural domains.

Statistical and Analytical Procedures

For quantitative analysis, repeated-measures ANOVA was used to evaluate within-subjects change across the three time points, supplemented by pairwise t-tests with Bonferroni correction. Effect sizes for repeated-measures analyses were reported as partial eta-squared (ηp^2) and interpreted using benchmarks commonly applied in L2 research (Plonsky & Oswald, 2014).

Repeated-measures ANOVA was conducted for complete data across time points. The assumption of sphericity was assessed using Mauchly's test; where violated, Greenhouse-Geisser corrections were applied. To examine differential effects by proficiency level, mixed-design ANOVAs were conducted with proficiency as a between-subjects factor and time as a within-subjects factor.

Separate repeated-measures ANOVAs were conducted for each error type to assess category-specific improvement using Bonferroni correction ($\alpha = .007$ for seven categories).

Of the 77 initially enrolled, 73 participants completed all three testing points (dropout rate: 5.2%). Four students from the intervention groups withdrew due to various personal reasons. The analysis of pre-test error rates did not reveal any systematic differences between completers and non-completers ($t = 0.34$, $p = .73$). All 13 comparison-group students completed the study. The required sample size was estimated through an a priori power analysis conducted using G*Power 3.1 (Faul et al., 2007) for a repeated-measures ANOVA design ($\alpha = .05$, power = .80, medium effect size f = .25). The analysis indicated a required sample of n = 28 per proficiency group to detect medium effects.

Qualitative analysis involved a focused review of AI feedback logs and student reflections submitted voluntarily. While these data were not subjected to formal coding procedures, they were analyzed to identify recurring patterns that complement the quantitative findings. These insights are presented as exploratory and supportive rather than as primary evidence.

Ethical Considerations

The study adhered to institutional ethical guidelines for research involving human participants. As the study involved non-invasive educational practices with voluntary participation and anonymized data, formal ethical review was not required. Informed consent was obtained from all participants prior to data collection.

Use of Generative AI

ChatGPT (OpenAI; GPT-4 and GPT-4 Turbo variants) was used in this study for instructional purposes and data processing. ChatGPT (OpenAI) was used to provide feedback on learner speech, and Whisper (OpenAI) was used for transcription. All AI-generated outputs were manually reviewed and verified by the researcher. AI tools were not used to generate research results or statistical analyses.

Results

This section presents findings related to overall grammatical accuracy development, category-specific changes, proficiency-level differences, and comparison group outcomes. Grammatical accuracy (measured as errors per 100 words) improved significantly across all proficiency groups over the 14-week intervention. Mean error rates decreased from 6.3–12.4 (pre-test) to 4.0–7.1 (post-test), representing reductions of 36.5%–50.9%. The comparison group showed minimal change (−8.3%), which was not statistically significant ($p = .41$), providing contextual contrast rather than a basis for causal comparison.

All proficiency groups demonstrated reductions in grammatical error rates across time, whereas the comparison group showed minimal improvement despite participating in similar once-weekly classroom activities.

Descriptive statistics are presented in Table 1.

Table 1. Descriptive Statistics of Grammatical Error Rates Across Three Time Points

Group	Pre-test	Mid-test	Post-test	Change (%)
Low proficiency	12.4	9.8	7.1	-42.7%
Moderate Proficiency	10.6	7.0	5.2	-50.9%
High proficiency	6.3	4.9	4.0	-36.5%
Comparison group	10.8	10.3	9.9	-8.3%

Note. Within-group comparisons were significant ($p < .001$). The comparison group change was not significant ($p = .41$).

Figure 1 illustrates how each category evolved over the three testing points among intervention-group students.

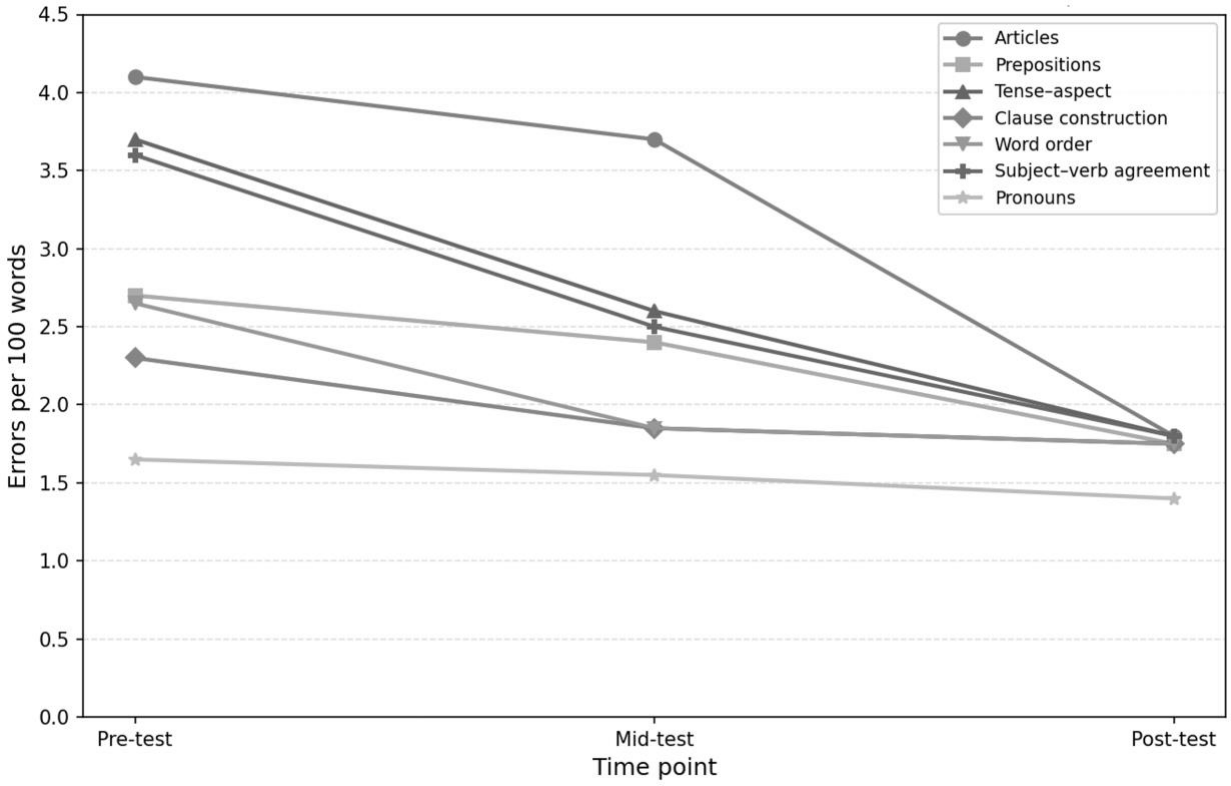

Figure 1. Trajectory of Error Reduction by Category Across Three Time Points

As shown in Figure 1, tense–aspect errors, subject–verb agreement errors, and basic word-order deviations exhibited the most consistent and steepest decline over time. Tense–aspect and subject–verb agreement errors showed particularly strong reductions across all three testing points. Error categories associated with greater cross-linguistic differences, such as articles and prepositions, demonstrated more gradual improvement. Although article errors decreased substantially over the intervention period, they remained among the most frequent error types at post-test. Clause-construction errors also declined gradually, whereas pronoun errors remained comparatively infrequent and did not reach statistical significance after Bonferroni correction.

Overall Grammatical Accuracy Development

A repeated-measures ANOVA revealed a significant main effect of time on grammatical accuracy ($F(2,118) = 51.83$, $p < .001$, $\eta p^2 = .45$). Given the non-randomized nature and small size of the comparison group (n = 13), it was not formally included in the repeated-measures ANOVA; instead, its outcomes are reported separately as a contextual reference rather than a basis for causal inference. In contrast, the comparison group showed minimal, non-significant change (−8.3%, $p = .41$).

Table 2. Overall Grammatical Error Rate Across Three Time Points

Condition	Mean	Standard Deviation
Pre-test	9.8	3.2
Mid-test	7.2	2.6
Post-test	5.4	2.1

Note: Sphericity verified by Mauchly's test (p = .14)

Bonferroni-adjusted pairwise comparisons show statistically significant improvement at all measurement points.

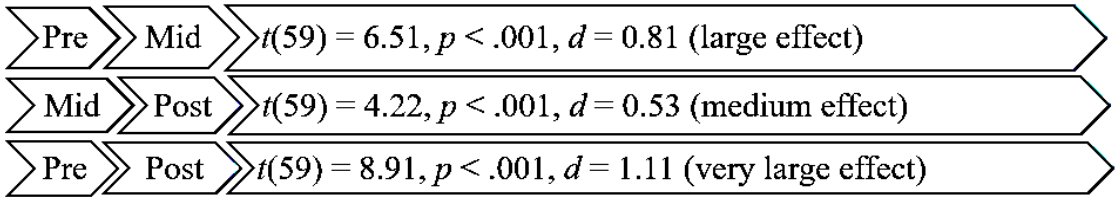

Figure 2. Bonferroni-Adjusted Pairwise Comparisons Between Time Points

A mixed-design ANOVA with proficiency as a between-subjects factor and time as a within-subjects factor indicates that the proficiency groups improved at different rates.

Table 3. Mixed design ANOVA: Proficiency Group x Time

Factor / Effect	F-statistic	p-value	Partial Eta Squared (ηp2)
Main effect of time *(Within-Subjects)*	F(2, 116) = 49.27	< .001	.46
Main effect of proficiency *(Between-Subjects)*	F(2, 58) = 42.38	< .001	.59
Time x Proficiency interaction	F(4, 116) = 3.82	= .006	.11

Post-hoc comparisons indicated that moderate-proficiency learners demonstrated the largest overall reduction in grammatical errors across the intervention period.

Category-Specific Statistical Analysis

Separate repeated-measures ANOVAs for each error category were conducted with Bonferroni correction ($\alpha = .007$).

Table 4. Statistical Significance of Improvement by Error Category

Error Category	F(2,122)	p	ηp²	Interpretation
Tense-aspect	48.23	< .001	.44	Large
Subject–verb agreement	41.17	< .001	.40	Large
Word order	38.94	< .001	.39	Large
Articles	12.43	< .001	.17	Medium
Prepositions	9.87	< .001	.14	Medium
Clause construction	15.32	< .001	.20	Medium
Pronouns	4.21	.021	.06	Small

Note. Separate repeated-measures ANOVAs were conducted for each error category (Bonferroni-adjusted $\alpha = .007$). Tense–aspect, subject–verb agreement, and word order showed large effect sizes ($\eta p^2 > .36$), whereas articles, prepositions, and clause construction demonstrated smaller but statistically significant effects. Pronoun errors did not reach statistical significance after Bonferroni correction, although a small trend-level effect was observed ($\eta p^2 = .06$). A two-way mixed ANOVA, including time and error category as within-subject factors, confirmed a significant time × category interaction, $F(12, 708) = 8.47$, $p < .001$, $\eta p^2 = .12$.

Comparison Case Study Outcomes

In the small, non-randomized comparison case study group ($n = 13$), overall grammatical accuracy remained relatively stable. Pre-test error rates ($M = 10.8$ per 100 words) decreased only marginally to post-test ($M = 9.9$). The decrease was not statistically significant ($p = .41$, $d = 0.18$). Individual-level analysis revealed that only 3 students (23%) demonstrated meaningful improvement (reduction > 1 SD) in persistent tense-aspect and agreement errors.

Qualitative Patterns in AI Feedback Engagement

Analysis of student feedback logs revealed several recurring patterns. First, learners demonstrated increased awareness of metalanguage, often anticipating and correcting habitual errors; many described AI feedback as functioning as "*mini grammar lessons*." Second, the persistence of AI memory was perceived as beneficial, as it reinforced awareness of recurring error patterns. Third, some learners reported overreliance on AI reformulations, occasionally adopting corrected outputs without engaging with underlying rules. Finally, engagement depth varied, with some learners actively interacting with feedback through follow-up questions, while others made only surface-level corrections. Moderate-proficiency learners appeared to engage more actively, although this observation remains exploratory.

Discussion

The findings of this study directly address the research questions and provide evidence for the role of structured AI-mediated feedback in grammatical development. The results indicate that structured pedagogical design combined with memory-enabled AI feedback is associated with the reduction of persistent grammatical errors in spoken production. Persistent errors in subject–verb agreement, tense-aspect, and word order were significantly reduced over the course of the intervention. These improvements align with findings that rule-governed L2 structures respond effectively to repeated corrective feedback (Nassaji, 2016). Memory-enabled AI appears to create conditions that support proceduralization, consistent with prior findings on corrective feedback (Li, 2010; Nassaji, 2016).

However, errors in articles, prepositions, and clause construction showed more gradual improvement. Armenian lacks articles as a grammatical category, requiring learners to develop new conceptual frameworks for definiteness and prepositional usage in the absence of L1 equivalents.

While many students developed metalinguistic awareness and self-correction strategies, others relied heavily on AI reformulations without deep processing. This divergence highlights the challenge of AI integration: balancing task-completion support with cognitive scaffolding.

Differential Responsiveness Across Proficiency Levels

The varying responsiveness across proficiency groups provides additional nuance. The most balanced and steady improvement was shown by learners with moderate proficiency, who had sufficient linguistic knowledge to comprehend AI explanations while still making enough errors for the feedback to bear significance.

Improvements among high-proficiency learners were less pronounced, likely due to the lower baseline error rates and the more subtle, discourse-level nature of their errors. The apparent lack of progress in the measured categories can be explained by the fact that those students started with lower error rates, indicating that they had already automated many fundamental structures.

Low-proficiency students, on the other hand, benefited during the intervention, although their trajectory was uneven. Cognitive overload may occasionally have resulted from metalinguistic explanations exceeding learners' processing capacity. These findings are consistent with the Zone of Proximal Development (ZPD) and with SLA research indicating that corrective feedback is most effective when aligned with learners' developmental readiness (Li, 2010).

Following benchmarks proposed by Plonsky and Oswald (2014), the observed effect sizes represent large effects in applied linguistics research. The large effect of time ($\eta p^2 = .45$) indicates substantial improvement in grammatical accuracy over the 14-week intervention. This magnitude exceeds those typically reported in corrective feedback research, where meta-analytic findings suggest more moderate effects (e.g., $d = 0.35$-0.58, approximately equivalent to $\eta p^2 = .03$-$.08$ in Li, 2010; $d = 0.40$-0.65, approximately $\eta p^2 = .04$-$.10$ in Nassaji, 2016). However, direct comparison should be interpreted cautiously due to differences in study design, duration, and outcome measures. The comparatively large effect observed here may be associated with the longitudinal and intensive nature of the intervention, including repeated exposure to individualized feedback, systematic tracking of persistent errors, and distributed practice across weekly cycles. In addition, the low-stakes, AI-mediated environment may have facilitated sustained engagement

with feedback. At the same time, variation in individual engagement levels, as practice time was guided but not strictly controlled, may also have contributed to the observed effect size.

Implications for Pedagogical Design

Students in the comparison case study group received similar speaking instruction without structured AI coaching, but did not demonstrate statistically significant improvement (M = 10.8 to M = 9.9 per 100 words, -8.3%). As noted, these results should be interpreted cautiously due to the non-randomized design and are illustrative rather than causal. The observed difference should not be interpreted as evidence that the limited improvement was caused by the absence of AI coaching.

The comparison group's experience of using AI represents a case in which students had access to AI without effective pedagogical guidance. Although these students frequently interacted with AI for content creation and memorization, their grammar did not improve as a result. The comparison group appeared to engage with AI in a predominantly product-oriented manner (e.g., content generation and memorization), although this was not systematically measured. This pattern aligns with a concern increasingly noted in recent research that task-completion-oriented AI use may be associated with reduced cognitive engagement (Abubakar et al., 2025).

The fact that only 3 of 13 students (23%) showed meaningful improvement (defined as ≥ 20% reduction in error rate), and that these three had independently sought additional practice, suggests intrinsic motivation and self-directed learning, rather than classroom instruction alone, were likely drivers of their progress.

These findings suggest that the effectiveness of AI tools depends largely on how they are pedagogically integrated. When used primarily for content generation and memorization, AI engagement does not appear to support measurable gains in grammatical accuracy, whereas the same system, when embedded in structured, feedback-oriented practice, is associated with substantial improvement.

Theoretical Implications for Skill Acquisition

Our findings support SAT's proposition that repeated, personalized feedback turns declarative knowledge into proceduralized performance (DeKeyser & Suzuki, 2025). The rapid initial gains followed by gradual refinement are similar to the learning progress during typical skill acquisition. However, differential effects across error categories indicate limitations of proceduralization: rule-governed structures showed large effects, while conceptually complex structures showed smaller effects. This distinction highlights the need for hybrid pedagogical models combining AI-driven micro-level feedback with human-led conceptual instruction. L1 typology constrains proceduralization through AI alone.

On Learner Autonomy and AI Overreliance

Balancing learner autonomy with AI support remains a critical issue. While AI-mediated feedback facilitated error awareness, some learners appeared to prioritize task completion over deeper processing, relying on AI reformulations without fully engaging with underlying grammatical principles. This pattern suggests that access to corrective feedback alone does not guarantee internalization, particularly when learners adopt efficiency-oriented strategies that may limit cognitive engagement.

Such tendencies may reflect a shift toward more passive learning behaviors in AI-supported environments, where efficiency can override sustained attention to form. As a result,

improvements in accuracy may remain surface-level rather than leading to durable competence. This underscores the importance of considering not only the availability of feedback but also the quality of learner engagement with it. Similar concerns have been noted in recent research on AI-supported learning, which highlights the risk of reduced cognitive engagement when learners rely excessively on automated outputs (Abubakar et al., 2025).

These findings suggest that AI feedback should augment rather than replace teacher-facilitated instruction. The comparison group's minimal improvement under unguided AI use indicates that neither AI nor traditional instruction alone is sufficient; their combination appears more effective. AI feedback logs can help teachers diagnose recurring errors and adjust instruction accordingly (e.g., targeting article usage based on learner data). Memory-enabled AI tools appear particularly effective for recurrent, rule-based errors (e.g., agreement, tense), allowing instructors to focus classroom time on more conceptually complex grammatical structures.

Instruction may be optimized when teachers focus on conceptual and discourse-level issues while AI provides repeated form-focused practice.

Educators should provide explicit instruction on how to interpret and apply AI-generated feedback. Without such guidance, some learners may memorize corrected outputs rather than engage with underlying grammatical principles. Instruction should therefore encourage active engagement with feedback, including requesting explanations, generating contrastive examples, and reflecting on rule application.

Incorporating guided analysis of anonymized AI feedback logs into classroom practice may further support deeper processing. By examining representative examples collectively, students can focus on generalizable patterns rather than isolated corrections, facilitating transfer to independent production.

AI-mediated speaking practice should also be integrated progressively, with task complexity increasing in alignment with learners' developing proficiency. A structured progression from simpler to more complex speaking tasks can support scaffolded development, particularly when combined with AI memory functions that reinforce attention to recurring errors over time.

This approach situates grammatical development within communicative use rather than treating it as an isolated skill, aligning with task-based language teaching principles and supporting transfer to spontaneous speech.

Implementation of hybrid feedback models should be included in the agenda of HEIs. In these models, AI tools assist with micro-level grammatical refinement while instructors concentrate on discourse-level accuracy, speech planning, and higher-order reasoning. This approach optimizes instructional effectiveness without sacrificing pedagogical integrity.

When combined with clear pedagogical guidance, AI tools provide a scalable solution for large-enrolment contexts where personalized feedback has traditionally been infeasible.

HEIs should develop institutional guidelines for the ethical use of memory-enabled AI systems, addressing openness, privacy, and the reduction of excessive reliance. Policies should foster transparent communication about AI's potential and limitations, promote its use as a practice tool rather than a content generator, and incorporate regular reflection exercises. Data retention protocols and assessment designs that measure internalized competence rather than AI-assisted performance are also needed.

These findings suggest that students benefit from being encouraged actively with AI to enhance their competence rather than use it as a source of ready-made outputs. The comparison group's transactional use of AI to generate content for memorization is a practice that should be avoided. To preserve learner agency, open communication about the purpose and limitations of AI is crucial.

Interpreting AI feedback logs and recognizing patterns across groups requires continuous teacher development strategies, the implementation of which will help instructors intervene pedagogically where AI is insufficient. This ensures the integration of AI literacy into teacher education curricula, which appears to be a growing priority in many educational systems (Daher, 2025).

- Designing effective AI prompts for language learning contexts;
- Analyzing AI feedback logs for diagnostic purposes;
- Balancing AI-mediated practice with human-led instruction;
- Recognizing and addressing signs of AI overreliance;
- Fostering critical evaluation of AI-generated content.

This capacity-building approach prepares teachers to leverage AI tools effectively; otherwise, students are left to navigate the pitfalls of AI use without guidance, a scenario that characterized the comparison case study group.

Thus, this study makes three main contributions:

- First, it provides longitudinal empirical evidence that structured, memory-enabled AI coaching improves grammatical accuracy in spontaneous spoken English.
- Second, it supports Skill Acquisition Theory in AI-mediated contexts, showing stronger effects for rule-governed structures than for conceptually complex categories.
- Third, it demonstrates methodological rigor through manual ASR verification, reducing measurement bias.

Limitations

Although the study provides valuable insights, a number of limitations should be acknowledged, and recommendations for future research should be made.

The study has limited generalizability, as it was conducted within a single institutional context and focused on one L1 background. As Armenian learners present a typological distance from English, the findings may vary for learners whose L1 shares more structural characteristics with English. To evaluate the generalizability of the findings, replication with students from different linguistic backgrounds is required.

The comparison case study group cannot support causal inferences due to variations in year levels, faculty composition, and AI engagement patterns, and the absence of random assignment. Future studies should use matched samples and randomized controlled designs.

Although AI feedback logs offered valuable contextual information, the analysis was an informal review dependent on students' voluntary submissions. Therefore, these qualitative findings should be viewed as illustrative rather than systematic or generalizable. To gain a more profound understanding of learners' cognitive processes, future research should incorporate more structured

qualitative techniques, such as think-aloud protocols during AI interactions or stimulated recall interviews.

There is a likelihood that the dual role of instructor and researcher might lead to observer bias during transcript verification. Moreover, students may alter their behavior to gain the instructor's favor. Although high inter-rater reliability (κ = .89) mitigates coding bias, future research should use coders blind to time points, proficiency levels, and study objectives.

The intensity of transcript verification represents another limitation. While manual transcript verification enhanced reliability, it was a labor-intensive process that might not be practical for extensive institutional research. The burden of verification would be reduced without sacrificing data quality if more precise ASR systems were developed and trained on learner speech from a variety of L1 backgrounds.

The AI tool used in this study was primarily text-based; therefore, feedback was provided through post-hoc transcript analysis rather than real-time speech processing. Future studies could examine how real-time AI feedback during speaking tasks might affect performance, possibly enabling instantaneous self-correction. Although Bonferroni correction was used to control Type I error, future studies may consider Holm's procedure, which is often less conservative for repeated comparisons.

The study did not thoroughly examine fluency or complexity, focusing instead on accuracy. Therefore, it remains unclear whether improvements in grammatical accuracy came at the expense of some students' fluency. A comprehensive CAF analysis would provide a more complete picture of oral development. Although Bonferroni correction was used to control Type I error, future studies may consider Holm's procedure, which is often less conservative for repeated comparisons.

Long-term retention was not measured. Although significant progress was observed during the 14 weeks, further research is needed to determine if learners continue to show improvements after the intervention ends. The durability of AI-mediated grammatical gains would be revealed by follow-up testing conducted at intervals of three and six months.

The AI model received multiple updates during the study period (from GPT-4 to GPT-4 Turbo variants). While memory functionality remained stable, subtle changes in feedback patterns may have occurred, potentially affecting reproducibility. Replication studies should document model versions clearly and use frozen API versions where feasible, as findings may not transfer to future AI systems with different capabilities.

The effectiveness of AI feedback may be moderated by individual difference variables, such as working memory capacity, aptitude, motivation, or prior language learning experience, but this was not systematically examined in the study. More focused implementation would result from knowing which students gain the most from memory-enabled AI coaching.

Conclusion and Future Directions

This study provides evidence suggesting that memory-enabled AI speech coaching is associated with reductions in L1-influenced grammatical errors in the spontaneous speech of Armenian EFL learners. The most pronounced improvements were observed in rule-governed structures, such as tense–aspect consistency, subject–verb agreement, and word order, whereas conceptually complex areas such as prepositions and articles showed more gradual progress. Moderate-proficiency

learners benefited most consistently, even though all intervention groups showed statistically significant and pedagogically substantial gains.

Persistent error tracking through AI memory appears to maintain students' attention on the recurring grammatical patterns, helping them transition from declarative to proceduralized competence. However, instructors should provide guidance to learners to prevent overreliance on AI as a universal problem-solving tool. The comparison group's minimal progress, despite having access to AI tools, underscores that learning outcomes are shaped by pedagogical design rather than technology alone, particularly in the absence of structured, feedback-oriented engagement.

AI use limited to transactional purposes, such as content generation for memorization, undermines long-term skill development. Iterative use of AI, including practice, feedback, and reflection within structured pedagogy, is associated with measurable gains.

The gradual incorporation of AI tools in Higher Education underscores the significance of our pedagogically informed research in explicating how these systems can best support sustainable development in L2 speaking. The findings suggest that memory-enabled AI should be regarded as a supplementary tool within structured speaking instruction, rather than as an independent instructional solution.

When integrated into structured curricula with explicit pedagogical direction, memory-enabled AI tools can facilitate grammatical advancement. However, such tools cannot replace the primary role of teachers, who deliver conceptual instruction, enhance metalinguistic awareness, and promote learners' critical engagement with AI-generated feedback. Consequently, AI's role in language education should be viewed as enhancing pedagogical practices rather than automating instruction.

Future Research Directions

Future research should examine several directions. First, longitudinal studies are needed to evaluate the durability of AI-mediated grammatical gains through follow-up assessments at three- and six-month intervals. Second, research comparing real-time AI feedback with post-task feedback could clarify optimal timing for proceduralization. Third, replication across learners with diverse L1 backgrounds would help determine the generalizability of findings and identify which grammatical features are most responsive to AI support. Fourth, experimental studies comparing AI-only, human-only, and hybrid instructional models could clarify the most effective pedagogical configurations. Finally, future work should investigate individual difference variables, such as working memory, motivation, and aptitude, as well as extend the analysis beyond sentence-level accuracy to include discourse and pragmatic competence.

About the Author

Lilit Sargsyan is an Associate Professor and PhD in Philology at Yerevan State University, Armenia, with 28 years of teaching experience and over 15 years as a YSU teacher trainer. A recipient of the YSU Teaching Excellence Award, her research focuses on AI integration in language education, student-centered learning, and critical thinking development. ORCID ID: 0009-0007-2862-2900

To Cite this Article

Sargsyan, L. (2026). Pedagogically structured AI feedback and grammatical accuracy in EFL speaking: A 14-week longitudinal study. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 30*(2). https://doi.org/10.55593/ej.30118int1

References

Abubakar, S., Jeilani, A., & Yusuf, M. (2025). The role of over-reliance on AI in the negative consequences of student learning: The moderating effects of ethical concerns and institutional policies. *Cogent Education, 12*, 2591503. https://doi.org/10.1080/2331186X.2025.2591503

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Daher, R. (2025). Integrating AI literacy into teacher education: A critical perspective paper. *Discover Artificial Intelligence, 5*, 217. https://doi.org/10.1007/s44163-025-00475-7

DeKeyser, R. M., & Suzuki, Y. (2025). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (4th ed., pp. 157–182). Routledge.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Dum-Tragut, J. (2009). *Armenian: Modern eastern Armenian* (London Oriental and African Language Library, Vol. 14). John Benjamins. https://doi.org/10.1075/loall.14

Ellis, R. (2017). Task-based language teaching. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 108–125). Routledge. https://doi.org/10.4324/9781315676968

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/BF03193146

Jaganov, T., Blake, J., Villegas, J., & Carr, N. (2025). Large language model-driven dynamic assessment of grammatical accuracy in English language learner writing. *IEEE Access, 13*, 151538–151550. https://doi.org/10.1109/ACCESS.2025.3603191

Jarvis, S. (2011). Conceptual transfer: Crosslinguistic effects in categorization and construal. *Bilingualism: Language and Cognition, 14*, 1–8. https://doi.org/10.1017/S1366728910000155

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon.

Liu, G. L., & Zhao, X. (2026). A scoping review of AI-mediated informal language learning: Mapping out the terrain and identifying future directions. *ReCALL, 38*, 111–130. https://doi.org/10.1017/S0958344025100359

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*, 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

LoCamato, M., & Munro, M. J. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL*, *35*(2), 189–205. https://doi.org/10.1017/S0958344022000067

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). John Benjamins.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition, 19*, 37–66. https://doi.org/10.1017/S0272263197001034

Mingyan, M., Noordin, N., & Razali, A. B. (2025). Improving EFL speaking performance among undergraduate students with an AI-powered mobile app in after-class assignments: An empirical investigation. *Humanities and Social Sciences Communications, 12*, 33. https://doi.org/10.1057/s41599-025-04688-0

Mizumoto, A. (2025). Automated analysis of common errors in L2 learner production: Prototype web application development. *Studies in Second Language Acquisition, 47*(3), 867–884. https://doi.org/10.1017/S0272263125100934

Nassaji, H. (2016). Interactional feedback in second language teaching and learning: A synthesis and analysis of current research. *Language Teaching Research, 20*, 535–562. https://doi.org/10.1177/1362168816644940

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524537

OpenAI. (2026). *ChatGPT memory features*. https://help.openai.com/en/articles/8590148-memory-in-chatgpt

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878–912. https://doi.org/10.1111/lang.12079

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing, 26*, 10–27. https://doi.org/10.1016/j.jslw.2014.09.003

Appendix A: Error Coding and Sample Size Justification

Article error: Omission before singular countable nouns; misuse of *the* in generic contexts; inappropriate zero article.
Pronoun error: Confusion between *he/she*; demonstrative pronoun errors (*this/these*).
Tense-aspect error: Inconsistent or inappropriate use of present/past/perfect/progressive forms with clear temporal adverbials.
Word order error: Non-canonical syntax influenced by L1 topic-fronting or Armenian clause structure.
Preposition error: Addition (*discuss about*), omission, or substitution (*depend from*).
Subject–verb agreement error: Number mismatch under cognitive load (e.g., *He study*).
Clause construction error: Malformed relative clauses; incomplete subordinate clauses.

Inter-rater reliability (25% of transcripts): $\kappa = .89$, 95% CI [.85, .93].

Sample Size Justification

Using G*Power 3.1.9.7 for a priori power analysis, the sample size was established (Faul et al., 2007). A repeated-measures ANOVA with three time points, $\alpha = .05$, power = .80, and a conservative estimate of medium effect size $f = .25$) required $n = 28$ per proficiency group based on effect sizes from similar corrective feedback interventions (Li, 2010: $d = 0.58$, equivalent to $\eta p^2 \approx .08$). While acknowledging that the low-proficiency group is underpowered for smaller effects, our achieved sample (Low: $n = 18$, Moderate: $n = 29$, High: $n = 17$, total $N = 64$) offers sufficient power (.82) to detect medium-to-large effects.

Appendix B: AI Speech Coach Configuration Specifications

To ensure consistency across participants and replicability of the intervention, the AI speech coach was configured to operate under the following constraints:

- **Error identification:**
  To identify grammatical deviations at the sentence and clause-level, the AI analyzed post-task transcripts of students' spoken responses, categorizing mistakes according to a predefined typology (e.g., tense–aspect, subject–verb agreement, articles, prepositions, word order) rather than providing comprehensive or impressionistic evaluations.
- **Memory tracking:**
  To facilitate longitudinal tracking and the generation of personalized reminders related to previously identified persistent errors, the AI stored abstracted representations of learners' recurrent grammatical error patterns (such as "frequent article omission" or "tense inconsistency in narrative contexts") across sessions.
- **Feedback constraints:**
  The learner could only receive (a) brief metalinguistic explanations, (b) localised corrective reformulations of their own utterances, and (c) targeted prompts that encouraged self-correction. The AI was specifically instructed not to introduce new content, rewrite entire responses, or model complex structures that weren't in the student's original production.

- **Pedagogical focus:**
  The system minimized construct-irrelevant variation by emphasizing accuracy-focused feedback that was in accordance with the study's research objectives and by not offering evaluative scores, fluency ratings, or content-based critiques.

Appendix C: Comparison Group Design Rationale

Given the limitations such as year-level and faculty differences**,** absence of formal proficiency assessment, and small size of the cohort, we considered three design alternatives:

1. excluding the comparison group entirely,
2. recruiting matched first-year students,
3. maintaining the current design with explicit framing as an illustrative case.

We selected option (3) because this cohort provides an example of outcomes when comparable speaking instruction takes place without organized AI scaffolding.