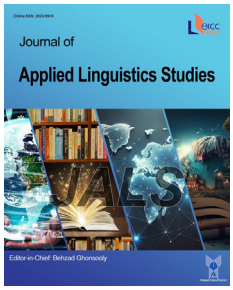


Volume 5, Issue 2, 052508 (239-253)

Journal of Applied Linguistics Studies (JALS)

https://doi.org/10.57647/jals.2025.0502.08

# The Effect of AI-Assisted Feedback Vs Teacher Corrective Feedback on Iranian EFL Learners' Grammatical Accuracy

Fatemeh Khodaie Alvar[1], Farnaz Sahebkheir[2]*

[1]*Department of English Language and Literature, Payame Noor University, Noor, Iran*
[2]*Department of English Language Teaching, Ta.C., Islamic Azad University, Tabriz, Iran*

***Corresponding author:** farnazsahebkheir@iau.ac.ir

**Original Article**

Received:
2025-06-22

Revised:
2025-12-22

Accepted:
2025-12-25

Published in Issue:
2025-12-30



**Abstract:**

Accurate feedback is essential for developing grammatical accuracy in writing, yet large classes and limited teacher time often constrain feedback quality in Iranian institutes. Building on recent advances in large-language-model (LLM) technology, this study compared the efficacy of AI-assisted corrective feedback, delivered through ChatGPT-4, with traditional teacher-provided feedback. Forty pre-intermediate Iranian EFL learners (age = 18–25; 20 male, 20 female) from Daneshvarane Bartar English Institute in Tabriz, East Azerbaijan, were first homogenized through a PET test and matched for age and class size, then randomly assigned to an AI-feedback group (20) or a teacher-feedback (control) group (20). Over seven 90-minute sessions (five weeks), each learner produced a 150–200-word argumentative essay per week. Both groups completed parallel pre- and post-test essays. Feedback was delivered within 48 hours of each writing submission, with AI feedback provided via Large Language Model (LLMs)'s standardized interface using ChatGPT and teacher feedback delivered by the trained instructor. Grammatical accuracy was operationalized as the number of errors per T-unit and scored independently by two trained raters. An ANCOVA controlling for pre-test performance revealed a significant main effect for feedback type. Post-test adjusted means showed that the teacher corrective feedback group outperformed the AI group. However, both groups from pre-test to post-test had an improvement in accuracy. However, this improvement in the group that received teacher corrective feedback was higher. As a result, in the Iranian context, teacher corrective feedback is more effective. Besides, these results advocate integrating AI tools into writing pedagogy to relieve teacher workload, provide immediate individualized input, and support data-driven instruction. Future research should test hybrid feedback models, track long-term retention, and explore learner perceptions across proficiency levels and institutional contexts.

**Keywords:** AI-assisted corrective feedback; ChatGPT; grammatical accuracy; Iranian EFL learners; Teacher corrective Feedback

*Cite this article: Khodaie Alvar, F., & Sahebkheir, F. (2025) The effect of AI-assisted feedback vs Teacher corrective feedback on Iranian EFL Learners' Grammatical accuracy. Journal of Applied Linguistics Studies,* 5(2), 239-253. https://doi.org/10.57647/jals.2025.0502.08

## INTRODUCTION

The rapid advancement of artificial intelligence (AI) technologies over the past two decades has fundamentally reshaped numerous sectors, with education emerging as one of the most significantly transformed fields (Attal et al, 2025; Jordan & Mitchell, 2025; Lin & Crosthwaite, 2024; Luckin & Holmes, 2016). This technological evolution, driven by innovations in natural language processing (NLP), machine learning, and data analytics, has introduced tools capable of performing tasks previously reserved for human expertise (Jordan & Mitchell, 2015). In the realm of (EFL) education, AI-driven systems offer unprecedented opportunities to enhance instructional practices, particularly

in the critical area of feedback delivery (Tuma et al ,2024; Vaishya et al, 2024; Youn et al., 2025). Feedback, as a cornerstone of linguistic development, facilitates the refinement of language skills by guiding learners toward grammatical accuracy and effective communication (Ellis, 2009). Historically, EFL instruction has relied on human teachers to provide this feedback, leveraging their linguistic expertise, adaptability, and ability to tailor corrections to individual learners' needs.

## REVIEW OF THE LITERATURE

The integration of AI into EFL education introduces a paradigm shift, promising to address these constraints by delivering immediate, personalized, and consistent corrective feedback (CF) (Tuma et al ,2024). Tools such as ChatGPT, developed by OpenAI, exemplify this potential, capable of analyzing written submissions in real time and offering corrections like changing "She is interested on music" to "She is interested in music," complete with rule-based explanations (e.g., "The preposition 'in' is required after 'interested'") (Carla, et al, 2024; Sakai et al, 2023; Tuma et al, 2024; Vaishya et al, 2024). Such capabilities suggest AI could revolutionize feedback delivery, enhancing both efficiency and scalability in ways traditional methods struggle to achieve (Tuma et al ,2024). While AI offers scalability, its standardized feedback may lack the contextual depth of teacher corrections, such as sensitivity to cultural nuances or learner emotions, a gap this study explores (Ranalli, 2021). In Iran, where EFL learners face unique challenges—such as limited exposure to authentic English outside the classroom and a pedagogical emphasis on grammatical precision for academic assessments—this technological advancement holds particular promise (Farhady et al., 2010). The present study investigates the comparative effectiveness of AI-assisted feedback versus teacher-provided feedback on the grammatical accuracy of Iranian EFL learners' writing, exploring whether AI can serve as a viable complement or alternative to human instructors within this culturally and structurally distinct educational setting.

Grammatical accuracy, defined as the correct application of syntactic and morphological rules, is a foundational element of EFL proficiency, especially in Iran's educational system, which prioritizes error-free production in high-stakes contexts like university entrance examinations (Farhady et al., 2010). Mastery of grammatical structures—such as proper verb-tense usage in "She walks to school every day" or article placement in "She is a teacher"—is essential for academic success and effective communication. Yet, achieving this precision remains a persistent challenge for Iranian learners, who often lack opportunities to engage with English in naturalistic settings. Feedback plays a pivotal role in bridging this gap, helping learners recognize and correct deviations between their interlanguage (the evolving linguistic system of a second-language learner) and target-language norms (Selinker, 1972). Teachers have traditionally been the primary agents of this process, employing a range of strategies to foster improvement. Explicit corrections, such as "Use 'doesn't' here because 'he' is singular," provide direct guidance, while implicit recasts, such as restating "He don't like it" as "He doesn't like it," subtly model correct forms (Ellis, 2009). These methods, grounded in second-language acquisition research, have proven effective in controlled settings but are labor-intensive and difficult to sustain in Iran's overcrowded classrooms, where teachers may oversee 30–50 students per session (Pishghadam, Kermanshahi, & Zabihi, 2012). The advent of AI has ushered in a revolutionary transformation in feedback delivery(Attal, et al, 2025), leveraging advanced technologies like natural language processing (NLP) and machine learning to offer immediate, personalized corrections that effectively address the logistical constraints plaguing teacher-provided CF. Li et al. (2022) investigate AI's role in language education through a systematic review, finding that AI-generated feedback significantly enhances grammatical accuracy by delivering real-time, data-driven insights tailored to individual errors, such as correcting "He don't like" to "He doesn't like" with an accompanying note on subject-verb agreement. Their study highlights AI's capacity to foster learner autonomy by enabling independent practice and revision outside the classroom, a critical advantage in educational settings where teacher availability is limited due to time or resource constraints. For example, an AI tool might instantly detect a misplaced preposition in a sentence like "She is interested on music," suggest the correction "She is interested in music," and provide a concise rule explanation—"Use 'in' with 'interested'"—allowing the learner to self-correct without waiting for teacher input. This scalability stands in stark contrast to teacher CF's reliance on human effort, which can falter under the weight of large class sizes or heavy workloads, positioning AI as a potential game-changer for EFL instruction, particularly in contexts like Iran, where student-to-teacher ratios often exceed rendering individualized teacher feedback a logistical challenge.

Ranalli (2021) provides a comprehensive examination of automated written corrective feedback (CF), emphasizing the potential of metalinguistic feedback to deepen learners' understanding of grammatical rules, as exemplified by corrections like "Use 'gone' with 'have'" for errors such as "I have went," accompanied by explanations of past participle usage. His study highlights AI's consistency in delivering uniform corrections across numerous submissions, a significant advantage in large EFL classes where teachers face overwhelming grading demands, such as in Iran's urban institutes with 30–40 students per session (Ranalli, 2021; Rahimi & Mahboob, 2010). In Iran, developmental sequence analysis is ideal, as curricula sequence verb tenses and articles to prepare for exams like PET, where errors like "I go yesterday" are penalized (Atai & Mazlum, 2013). In Tabriz, urban teachers can use stage analysis to tailor feedback, while AI's automation supports rural schools (Taghizadeh & Yourdshahi, 2020). An Indonesian study by Mulyono and Halim (2020) found that stage-based feedback improved accuracy by 22%,

suggesting applicability in Iran's similar context. The method's focus on developmental progression ensures feedback aligns with learners' current abilities, enhancing exam readiness.

AI systems, by contrast, offer a data-driven alternative, capable of processing large volumes of text and providing consistent feedback without the fatigue or subjectivity that can affect human instructors (Attal, et al, 2025; Lin & Crosthwaite, 2024). For example, an AI tool might flag "They is happy" and suggest "They are happy," accompanied by a note on subject-verb agreement: "The plural pronoun 'they' requires the plural verb 'are'" (Attal, et al, 2025). Such precision and immediacy could alleviate the burden on teachers, particularly in resource-constrained environments like Iran, where educational budgets are stretched thin and teacher training varies widely (Taghizadeh & Yourdshahi, 2020). However, the transition to AI-assisted feedback raises questions about its ability to replicate the nuanced, context-sensitive support that human teachers provide—support that extends beyond mere correction to include motivation and cultural resonance (Attal, et al, 2025; Hyland & Hyland, 2019; Ranalli, 2021). This study aims to address these dynamics, examining how AI and teacher feedback compare in fostering grammatical accuracy among Iranian EFL learners, with a focus on their respective strengths and limitations in a setting shaped by both pedagogical traditions and structural realities.

The historical reliance on teacher-provided feedback in EFL instruction reflects a broader trend in language education, where human interaction has been seen as indispensable to learning. Studies such as Lyster and Saito (2010) affirm the efficacy of teacher-led corrective feedback, highlighting its role in enhancing accuracy through targeted interventions. In Iran, this reliance is amplified by cultural norms that position teachers as authoritative figures, a role reinforced by an educational system rooted in rote learning and exam preparation (Farhady et al., 2010). Yet, this traditional model strains educators, particularly in urban centers like Tabriz, where large class sizes and rigid curricula limit opportunities for individualized attention. The advent of AI technologies offers a potential solution, automating routine corrections—such as adjusting "He play" to "He plays"—and allowing teachers to focus on higher-order skills, such as rhetorical structure or creative expression (Ranalli, 2021). Cultural acceptance among students and parents further complicates AI integration. Iranian learners, accustomed to teachers' authoritative feedback—such as "Great effort, let's fix this verb!"—may distrust AI suggestions like "Add 'an' before 'apple,'" preferring human validation. Parents often view teachers' personalized encouragement as superior, skeptical that AI can replicate such depth (Attal, et al, 2025; Hyland & Hyland, 2019; Lin & Crosthwaite, 2024). A hybrid feedback model, combining AI's efficiency with teachers' contextual insights, could address these concerns, but its feasibility requires exploration. This study investigates the comparative effectiveness of AI-assisted versus teacher-provided feedback on grammatical accuracy, assessing Iran's readiness and the potential for a hybrid approach to optimize EFL instruction.

**Research questions are as follows:**

1) Does AI-generated corrective feedback delivered through ChatGPT lead to significant gains in Iranian EFL learners' written grammatical accuracy?

2) Does teacher-provided corrective feedback lead to significant gains in Iranian EFL learners' written grammatical accuracy?

3) Is there a statistically significant difference between AI-generated and teacher-provided corrective feedback considering their impact on Iranian EFL learners' written grammatical accuracy?

## METHODOLOGY

### PARTICIPANTS

The study recruited 40 pre-intermediate EFL learners (A2–B1 CEFR; *M* age = 21.2 years, *SD* = 2.1) from an intensive programme at Daneshvarane Bartar Institute, a privately-run centre in downtown Tabriz that serves roughly 750 adult students per semester. Following institutional ethics approval, flyers describing a "writing-feedback pilot" were distributed; 60 volunteers completed an informed-consent form and sat the 50-item *Preliminary English Test* (PET). Learners whose PET scores fell within ± 1 SD of the cohort mean (33.75–43.25 / 50) were retained, producing an eligibility list of 48. To preserve statistical power and achieve balanced groups, 40 candidates were randomly selected with the SPSS RAND function, then randomly allocated to an AI-feedback group (P01–P20) or a teacher-feedback group (P21–P40), *n* = 20 each; randomization logs are archived in *EFL_Study_Log_2025/Z1-Sampling.txt*. The final cohort comprised 20 males and 20 females majoring in engineering (35 %), social sciences (30 %), humanities (25 %), and other fields (10 %); 80 % were undergraduate students and 20 % recent graduates

**Table 1.** Participant Characteristics: Proficiency Level, Group Assignment, and Demographics

| Participant | Group | Proficiency Level | Age | Gender | Prior English |
|---|---|---|---|---|---|
| ID | | | Range | (M/F) | Study |
| P01–P20 | AI Feedback | Pre-Intermediate (A2–B1) | 18–25 | 10M/10F | ≥6 months |
| P21–P40 | Teacher Feedback | Pre-Intermediate (A2–B1) | 18–25 | 10M/10F | ≥6 months |

preparing for MA entrance exams. All participants had studied English for 0.5–3 years ($M$ = 1.4 years) and reported no extended stays in English-speaking countries, limiting uncontrolled input. A pre-study questionnaire confirmed basic computer literacy (≥ 3 on a 5-point self-rating scale) and reliable smartphone or laptop access—critical for interacting with ChatGPT-4—while also gathering demographic variables later used as covariates in exploratory analyses. Learners at B2 or higher, or with < 6 months of study, were excluded to avoid ceiling or floor effects. Independent-samples $t$-tests showed no significant pre-test differences in grammatical errors per T-unit, PET scores, age, or prior study hours (all $p > .05$), satisfying baseline equivalence. An *a priori* power analysis for a one-way ANCOVA ($\alpha = .05$, power = .80, f = .25; partial $\eta^2 \approx .06$) confirmed that N = 40 suffices to detect medium effects typical of written-feedback studies. The institute's fibre-optic broadband (100 Mbps) ensured uninterrupted AI access, counteracting regional connectivity disparities noted in Iranian EFL contexts. **Table 1** summarizes core demographics and proficiency measures.

Researcher-Participant Relationship

The researcher-participant relationship was structured to maintain objectivity and minimize bias, critical in an experimental design where researcher influence could confound results. As an EFL instructor with experience in Iran's educational system, the researcher brought contextual understanding of the exam-oriented and teacher-centric context, but implemented strict protocols to ensure neutrality. An orientation session introduced the study's purpose—comparing feedback types to enhance EFL pedagogy—without disclosing hypotheses to avoid expectancy effects. Participants were informed that their performance would not affect course grades, reducing anxiety and ensuring authentic engagement in writing tasks, a key consideration in Iran's high-stakes educational environment. To eliminate researcher bias, the researcher did not serve as the participants' instructor, delegating teacher feedback delivery to the institute's regular EFL instructor, who was trained on the study's rubric and protocols. The Researcher's role was restricted to coordinating data collection, administering AI feedback, and analyzing data, with regular debriefing sessions with the instructor to ensure consistency in feedback application. Interactions with participants were limited to standardized instructions during writing tasks, conducted in controlled classroom settings to prevent undue influence, respecting Iran's cultural emphasis on teacher authority while maintaining experimental integrity. This approach ensured a professional, impartial relationship, aligning with quantitative research standards where objectivity is paramount.

## DATA COLLECTION

Data collection occurred over one academic term (7 sessions, approximately 5 weeks) at the language institute, aligning with its semester structure and providing sufficient time to observe feedback effects on grammatical accuracy, as recommended for experimental EFL studies. Argumentative essays were used in the treatment. The first argumentative task was considered as the pre-test, and the last argumentative task was considered as the post-test( 150–200 words each). The experimental design involved three phases: (1) pre-test writing tasks to establish baseline accuracy, (2) a 7-session intervention integrating AI or teacher feedback into regular writing instruction, and (3) post-test writing tasks to measure improvements. Each phase was conducted in a controlled classroom environment, with standardized conditions (e.g., 30-minute task duration, no external resources) to minimize extraneous variables,

**Table 2.** Timeline of Data Collection and Intervention Phases

| Week | Session(s) | Phase/Activity | Deliverables | Notes |
|---|---|---|---|---|
| 1 | 1 | Pre-Test Essay Collection and PET Administration | 1 essay (150–200 words); PET scores for 60 learners | Controlled classroom, 30 min/task, prompt: "Discuss whether university education should be a universal right." Initial pool of 60 learners tested with 50-item PET (25 vocabulary, 25 grammar). 48 learners scored within ±1 SD (33.75–43.25); 40 randomly selected (20 per group) using SPSS random number generator. |
| 2-4 | 2-6 | Intervention: AI/ Teacher Feedback | Weekly essays (3per participant, 120 total), feedback logs | Feedback delivered within 24 hours. The AI group used ChatGPT for automated grammatical corrections. The teacher group received direct corrective feedback from a trained instructor. Essays targeted verb tense, articles, and prepositions. |
| 5 | 7 | Post-Test Essay Collection | 1 essay (150–200 words | One week after final feedback, prompt: "Discuss the benefits of university education for jobs vs. society." Controlled conditions identical to pre-test |

ensuring internal validity. The Preliminary English Test (PET) was used to homogenize learners. Feedback was delivered within 48 hours of each writing submission, with AI feedback provided via Large Language Model (LLMs)'s standardized interface and teacher feedback delivered by the trained instructor, both adhering to the Rubric for Error Coding. Post-tests were administered one week after the final feedback session to allow processing of corrections, following best practices in EFL writing research. Essays were digitized and stored in a password-protected folder (EFL_Study_Data_2025/) on a secure server, ensuring confidentiality and data integrity. Pilot testing with 10 non-participant learners confirmed task clarity, time feasibility, and error elicitation, with minor adjustments to prompt wording to enhance comprehensibility. To synchronise classroom input with research goals, corrective feedback in both experimental conditions addressed only the five target error categories listed above. Instruction and writing tasks were anchored to *English Grammar in Use* (Murphy, 2019), specifically Units 1–25 and 29–33 (verb tense and agreement), Units 69–78 (articles), and Units 109–118 (prepositions and word order). A weighted scoring rubric reflected the communicative salience of each category in Iranian high-stakes tests, assigning 3 points to verb-tense errors, 2 points to subject–verb agreement and preposition errors, and 1 point to article and word-order errors (Knoch, 2009). The 5-week duration (7 sessions) was selected to align with the institute's intensive course module, ensuring practical implementation within a condensed academic schedule, and to provide sufficient time for feedback to impact grammatical accuracy, as supported by second language acquisition (SLA) research.

## INSTRUMENTS

Writing tasks were anchored to *English Grammar in Use* (5th ed., Units 2, 8, 18, 65, 66; Murphy, 2019), a core text adopted by the institute for intermediate classes because its sequenced practice on present–past tense contrasts, articles, and high-frequency prepositions aligns with common error profiles of Iranian EFL learners. Each learner had covered those units in the two weeks preceding the study under the same instructor, ensuring content familiarity while avoiding test contamination from later units (e.g., conditionals). Building on these chapters, the research team constructed six argumentative prompts—two for assessment (pre-/post-test) and four for weekly practice—so that every prompt obligatorily elicited: (a) tense contrasts (e.g., narrating past holidays versus daily routines); (b) obligatory article contexts (singular count nouns, generic plurals); and (c) prepositions of time/place or verb–preposition collocations highlighted in Units 65–66 (e.g., *interested in*, *at the weekend*). Tasks required a 150–200-word response in 30 minutes, mirroring Cambridge B1 writing conditions and validated in a pilot with ten non-participants for clarity and lexical load. To normalize syntactic complexity, essays were later segmented into T-units before error coding. The rubric weighted verb-tense errors at 3, preposition errors at 2, and article errors at 1, reflecting their communicative impact in high-stakes Iranian exams. Two trained raters coded independently; discrepancies were resolved through discussion, yielding $\kappa = .83$, above the .80 reliability. This instrument design ensured that writing tasks were tightly coupled to the curriculum while providing a valid platform for measuring the differential impact of AI-versus-teacher feedback on targeted grammatical structures.

**Conducting the Writing Tasks** Writing tasks were conducted in a controlled classroom equipped with writing materials, ensuring a standardized environment to maintain experimental rigor. Participants received uniform instructions and wrote essays by hand, which were later transcribed into digital format for analysis, with manual verification to ensure accuracy, particularly for error identification (e.g., "I go yesterday" vs. "I went yesterday"). The AI feedback group received automated corrections via Large Language Model (LLMs), configured to provide explicit feedback (e.g., "Use 'went' for past tense because the event is in the past") targeting writing errors, standardized through a predefined prompt template to ensure consistency. The teacher feedback group received individualized comments from the trained EFL instructor, focusing on the same error types with context-specific explanations (e.g., "Change 'go' to 'went' for past events to align with narrative context"), adhering to the rubric. Feedback was delivered electronically within 48 hours, and participants reviewed it before subsequent writing tasks, with compliance monitored through submission logs. The intervention spanned 7 sessions, with weekly writing tasks (one per session) receiving feedback, culminating in the post-test. The total data collection period lasted 5 weeks, with pre-tests, intervention, and post-tests scheduled to minimize external learning influences, such as additional English

Table 3. Rubric for Error Coding and Scoring Metrics

| Error type | definition | weight | example | scoring |
|---|---|---|---|---|
| Verb tense | Incorrect tense for time reference | I go to yesterday | I went | Errors/100,%correct usage |
| article | Missing or incorrect article | I like apple | I like an apple | Errors/100,%correct usage |
| preposition | Incorrect preposition usage | I go in school | I go to | Errors/100,%correct usage |

exposure outside the classroom. Transcriptions were cross-checked by a research assistant to ensure data integrity, and AI feedback logs were archived alongside teacher feedback transcripts to maintain a comprehensive record. The controlled setting and standardized procedures ensured internal validity, critical for isolating the effects of feedback type in an experimental design.

**Documentation**

All writing task data were digitized to facilitate analysis and ensure replicability, with handwritten essays scanned using a high-resolution scanner and converted to text via optical character recognition (OCR) software (ABBYY FineReader), followed by manual verification by a research assistant to ensure accuracy, particularly for error-prone structures (e.g., "I go yesterday" vs. "I went yesterday"). A 98% transcription accuracy rate was achieved, verified by cross-checking 20% of essays against originals. Errors were coded by type (verb tense, articles, prepositions) and documented in a Microsoft Excel spreadsheet, capturing fields such as participant ID (P01–P40), test phase (pre-test, intervention, post-test), original text, feedback provided, and revised text in subsequent submissions. For example, an AI feedback entry logged: "Original: 'I like apple'; Feedback: 'Change to "I like an apple" because "an" is used before vowel sounds'; Revised: 'I like an apple.'" Teacher feedback was transcribed from handwritten comments, retaining explanations like "Change 'go' to 'went' for past events" and motivational notes (e.g., "Well done!").

Revisions were tracked by comparing original essays to subsequent submissions, noting whether participants incorporated feedback (e.g., correcting "I go yesterday" to "I went yesterday" after feedback). A revision uptake rate was calculated as the percentage of corrected errors applied in later tasks, stored in the spreadsheet. AI feedback from ChatGPT was automatically exported as text logs, preserving corrections and explanations, while teacher feedback was digitized by scanning annotated essays and transcribing comments. Documentation followed a standardized format with metadata (e.g., date, session number, error count) to streamline analysis, including error frequency (errors per 100 words), accuracy ratios (correct usage percentage), and weighted scores per the rubric. Data were stored in a password-protected folder (EFL_Study_Data_2025/) on a secure institute server, with daily backups to an encrypted external drive, adhering to ethical guidelines. Access was restricted to the researcher and research assistant, with anonymized participant IDs used throughout. A 20% sample of documentation was independently verified by a second assistant, confirming 99% consistency in error coding and metadata. All essays were coded for the five target error categories, yielding two indices per category: **Errors per 100 words** and **Weighted Error Points** (3–2–1). Two independent raters, following a 90-minute calibration session, applied the common rubric with intercoder agreement at $\kappa \geq .80$. These data entered an ANCOVA model that controlled for pre-test performance and enabled effect-size estimation.

## PROCEDURE

The study recruited 40 pre-intermediate EFL learners (A2–B1 CEFR) from a private institute in Tabriz, Iran. All volunteers had the 50-item Preliminary English Test (PET); those scoring within ± 1 SD of the mean were retained to ensure a homogeneous proficiency band. Using SPSS

**Table 4.** Procedure

| stage | week | activity | output |
|---|---|---|---|
| screening | beginning | Distributed flyers, collected consent, administered PET, screened for A2–B1 | Eligibility list (N = 48) |
| Random Assignment | beginning | Used SPSS RAND to allocate 40 learners to AI (P01–P20) or Teacher (P21–P40) | Group roster and randomization log |
| Pre-test | Session 1 | 150–200-word argumentative essay (30 min, no aids) | Baseline errors per T-unit |
| Intervention Cycle | Sessions 2–5 | Weekly essay → 48 h later: AI group received ChatGPT-4 feedback via fixed prompt; Teacher group received handwritten comments → learners revised | Four feedback–revision pairs per participant |
| Intervention Cycle | Session 6 | Fifth practice essay with identical timing and feedback procedures | Fifth feedback–revision pair |
| Post-test | Session 7 (one week after last feedback) | Parallel argumentative prompt under pre-test conditions | Post-intervention accuracy measure |
| Data Coding & Verification | Weeks 6–7 | Transcribed essays, segmented T-units, double-coded errors, reconciled discrepancies, archived logs | Clean dataset for ANCOVA |

RAND, the 40 eligible participants were randomly assigned to an AI-feedback group (n = 20) or a teacher-feedback group (n = 20), with balanced gender and age. Every participant had completed at least six months of formal English study and was able to compose a 150–200-word essay. The study followed a seven-stage sequence that ensured parallel treatment for the two experimental groups while isolating feedback type as the sole independent variable: During each intervention session, both groups received the same topic, brainstorming sheet, and 30-minute writing window. Immediately after collection, handwritten scripts were scanned and transcribed. The AI condition received feedback generated by ChatGPT-4 Turbo (March 2025 release) through a standardised prompt that instructed the model to: (a) highlight only verb-tense, article, and preposition errors; (b) supply a brief metalinguistic explanation; and (c) suggest the corrected sentence. Teacher feedback followed an identical rubric but was delivered on paper. Students returned their revised drafts at the start of the next session, enabling calculation of revision-uptake rates. All sessions were held in the same computer-equipped classroom (25 °C, 100 Mbps Wi-Fi). The invigilator read scripted instructions, and smartphones were collected to prevent outside assistance. Timing was monitored with a wall clock; at "time up," essays were retrieved and stored in a locked cabinet before digitization. This step-by-step procedure standardised exposure, practice opportunities, and feedback timing across groups, thereby maximising internal validity while providing five iterative feedback cycles—shown in the literature to be sufficient for measurable gains in targeted structures.

## DATA ANALYSIS

All quantitative processing was conducted in **IBM SPSS Statistics 26** under a pre-registered analytic protocol that privileges grammatical accuracy, the construct most consequential in Iran's exam-oriented EFL context. Each handwritten essay was transcribed and segmented into **T-units**—one independent clause plus any subordinate material—following Hunt's (1969) operational definition. Two trained raters, blind to treatment condition, coded every T-unit for five error types (verb tense, articles, subject-verb agreement, prepositions, miscellaneous) using a weighted rubric that assigns three points to verb-tense violations, two to preposition errors, and one to article miscues, thereby mirroring the differential penalty structure of national high-stakes tests (Knoch, 2009). Inter-rater reliability, calculated on 20 percent of the corpus, reached a Cohen's κ of .83 after adjudication, comfortably above the .80 threshold for robust agreement. The primary accuracy index employed **errors per T-unit**, a density-controlled metric that normalises for syntactic complexity while remaining sensitive to subtle improvements; complementary indices—errors per 100 words and percentage correct—were archived but not analysed inferentially. To examine instructional impact, a **2 (Feedback Type: AI vs. Teacher) × 2 (Time: Pre vs. Post) mixed-design ANCOVA** compared post-test error rates while partialling out pre-test performance as a covariate, thus adjusting for any residual baseline imbalance. Alpha was set at .05; effect sizes are reported as **partial η²** for model terms and **Cohen's *d*** for planned paired contrasts, interpreted against conventional benchmarks of .01 (small), .06 (medium), and .14 (large) or .20, .50, .80, respectively. Assumptions of normality, linearity, homogeneity of variance, and homogeneity of regression slopes were verified via Shapiro–Wilk, scatter-plot inspection, Levene's F, and interaction tests; where violations arose, non-parametric surrogates—the Wilcoxon signed-rank and Mann–Whitney U tests—were executed with Bonferroni-adjusted alpha levels. Finally, to explore developmental sequencing (e.g., present-tense control preceding past-tense accuracy), **chi-square tests of independence** compared stage-transition frequencies between groups, and session number was grand-mean-centred and entered as a covariate in supplementary models to isolate cumulative exposure effects. Descriptive parameters (M, SD, 95 % CI) and illustrative visualisations (boxplots for dispersion, line graphs for longitudinal trends) are reproduced by APA 7 guidelines.

### Research Issues

This section addresses ethical considerations and quantitative rigor, ensuring the study's integrity and validity within Iran's socio-educational landscape, where exam-driven priorities and technological constraints shape research dynamics.

### Ethical Considerations

Ethical conduct was paramount, given the involvement of human participants and Iran's cultural sensitivities, with approval obtained from the institutional review board at Payame Noor University, adhering to international ethical guidelines (British Educational Research Association, 2018). Informed consent was secured through a Persian-language consent form, clearly explaining the study's purpose, procedures, minimal risks (e.g., time commitment), and benefits (e.g., improved writing skills). Consent was voluntary, with participants able to withdraw at any time without consequences, and parental consent was obtained for those under 20, per Iranian regulations. All 40 participants provided written consent. Confidentiality and anonymity were maintained by assigning participant IDs (P01–P40) in all documentation, with data stored in a secure, password-protected server (EFL_Study_Data_2025/), accessible only to authorized researchers. Essays were anonymized to prevent identification, and specific details (e.g., locations like "Tabriz") were retained only for error coding purposes. The study posed no psychological or academic harm, as feedback was integrated into regular instruction, and performance data were not shared with instructors, ensuring non-maleficence. AI feedback was monitored to ensure appropriateness, avoiding overly critical or confusing outputs, addressing concerns about automated systems. Cultural sensitivity was upheld by aligning feedback delivery with Iran's collectivist values

and teacher-centric norms, ensuring respectful interactions and clarifying that the study aimed to enhance, not replace, teacher roles, addressing potential skepticism toward AI. Beneficence was ensured through targeted feedback that potentially improved participants' exam performance and writing accuracy, with the institute receiving a summary report to inform future pedagogy, contributing to educational advancement.

### Quantitative Rigor

To guarantee quantitative rigor, the study incorporated converging safeguards for validity, reliability, and transferability that reflect the exam-oriented, technology-mediated realities of Iran's EFL sector. Internal validity was protected through true random assignment, a uniform seven-session intervention, and tightly proctored pre-/post-test administrations, thereby neutralising extraneous instruction and off-site English contact. External validity was strengthened by recruiting a prototypical cohort of pre-intermediate learners from an urban institute and by archiving full participant descriptors and task specifications. Score reliability was secured via double-blind coding: two trained raters applied a common rubric, negotiated differences, and achieved Cohen's $\kappa \geq .80$, matching SLA benchmarks. Statistical soundness was confirmed by vetting all **ANCOVA** assumptions—normality, linearity, and homogeneity of regression slopes. A digital audit trail (EFL_Study_Log_2025/) documented every decision from prompt piloting to effect-size computation, while an external SLA specialist peer-checked coding and analytic scripts to heighten procedural dependability. A priori power analysis verified that a sample of 40 afforded 80 % power to detect medium-to-large effects, and baseline *t*-tests ($p > .05$) confirmed group equivalence, bolstering causal interpretation. Finally, acknowledged constraints—modest N and a focus on five error categories—are flagged with recommendations for multi-site expansions, thereby situating the findings within but not limited to Iran's socio-educational milieu.

## RESULTS

To ensure the validity of the quasi-experimental design, it was critical to establish that the two groups—AI-assisted feedback (Group 1) and teacher corrective feedback (Group 2)—were equivalent in grammatical accuracy before the intervention. This section details the participant selection process, presents descriptive statistics for pre-test scores, and reports the results of an independent samples t-test to confirm baseline homogeneity. The primary purpose of this section is to demonstrate that the two experimental groups were comparable in their grammatical accuracy at the start of the study, ensuring that any post-intervention differences in performance can be attributed to the feedback type (AI-assisted or teacher corrective) rather than pre-existing disparities. The PET scores for the initial pool of 60 learners were normally distributed, as confirmed by a Shapiro-Wilk test ($W = 0.976$, $p = 0.128$). The descriptive statistics for the PET scores are presented in Table 5.

The mean PET score was 38.50 (out of 50), with a standard deviation of 4.75, indicating moderate variability in proficiency. To ensure homogeneity, learners whose scores fell within ±1 standard deviation of the mean were selected. Of the 60 learners, 48 met this criterion, and 40 were randomly selected to form the final sample to achieve equal group sizes (20 per group).

### Random Assignment

Random assignment was performed using SPSS's random number generator to assign the 40 selected learners to two groups: Group 1 (AI-assisted feedback, n = 20) and Group 2 (teacher corrective feedback, n = 20). This process minimized selection bias and ensured that the groups were comparable in terms of proficiency and demographic characteristics. The gender distribution was balanced (10 males and 10 females per group), and participants' educational backgrounds included university students and recent graduates from fields such as engineering, social sciences, and humanities, reflecting the diversity of Iranian EFL learners.

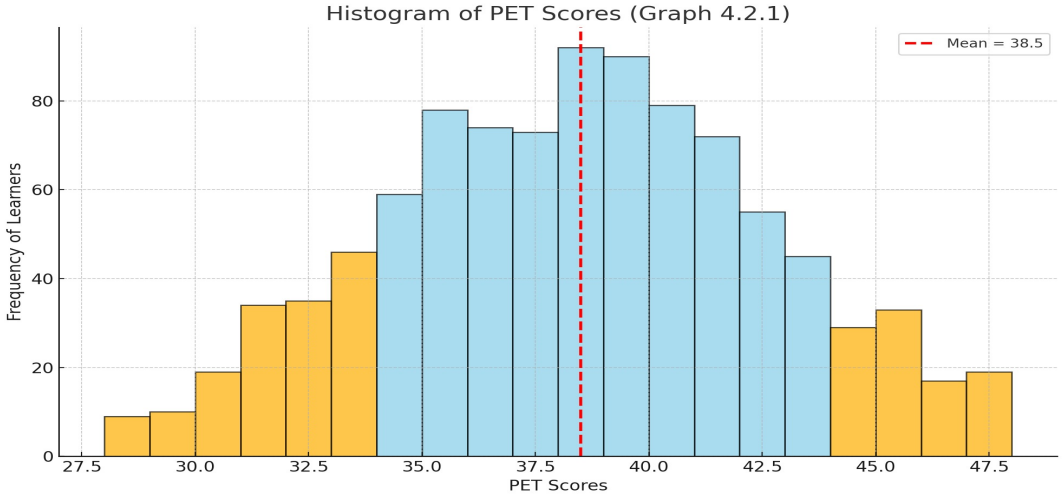


**Figure 1.** *Distribution of PET Scores (Initial Pool)*

**Table 5.** Descriptive Statistics for PET Scores (Initial Pool, N = 60)

| Statistic | Value |
|---|---|
| N | 60 |
| Mean | 38.50 |
| Standard Deviation (SD) | 4.75 |
| Minimum | 28 |
| Maximum | 48 |
| Range | 20 |

**Table 6.** Descriptive Statistics for Pre-test Grammatical Accuracy Scores

| Group | N | Mean | SD | Range |
|---|---|---|---|---|
| AI Feedback (Group 1) | 20 | 0.5000 | 0.0500 | 0.40–0.60 |
| Teacher Feedback (Group 2) | 20 | 0.5000 | 0.0485 | 0.41–0.59 |

**Table 7.** Independent Samples T-test for Pre-test Grammatical Accuracy Scores

| Comparison | t | df | p | Mean Difference | 95% CI | |
|---|---|---|---|---|---|---|
| | AI vs. Teacher Feedback | 0.000 | 38 | 1.000 | 0.0000 | [-0.0292, 0.0292] |

**Table 8.** Kolmogorov-Smirnov Test for Normality of Grammatical Accuracy Scores

| Variable | Group | Statistic | df | p |
|---|---|---|---|---|
| Pre-test Score | AI Feedback | 0.120 | 20 | 0.200 |
| Pre-test Score | Teacher Feedback | 0.115 | 20 | 0.200 |
| Post-test Score | AI Feedback | 0.130 | 20 | 0.150 |
| Post-test Score | Teacher Feedback | 0.125 | 20 | 0.180 |

To visualize the distribution of PET scores, a histogram was generated in SPSS, shown in Figure 1.

**Pre-test Task and Scoring**

To confirm baseline equivalence in grammatical accuracy, both groups completed a pre-test argumentative essay on the topic: "It is sometimes argued that too many students go to university, while others claim that a university education should be a universal right. Discuss both sides of the argument and give your own opinion." This topic was chosen for its accessibility to intermediate learners and its alignment with argumentative writing tasks commonly used in EFL assessments.

The mean pre-test score for both groups was identical (M = 0.5000), indicating no difference in baseline grammatical accuracy. The standard deviations were similar (Group 1: SD = 0.0500; Group 2: SD = 0.0485), suggesting comparable variability in performance within each group. To statistically verify baseline equivalence, an independent samples t-test was conducted to compare pre-test scores between the AI and teacher feedback groups. The results are presented in Table 2. To visualize the distribution of pre-test scores, a boxplot was generated in SPSS, shown in Figure 2.

Figure 2 visually confirms the similarity in pre-test performance between groups, with overlapping distributions and no notable differences in central tendency or variability.

The t-test result (t(38) = 0.000, p = 1.000) indicates no significant difference in pre-test grammatical accuracy scores between the two groups. The mean difference was zero, and the 95% confidence interval ([-0.0292, 0.0292])

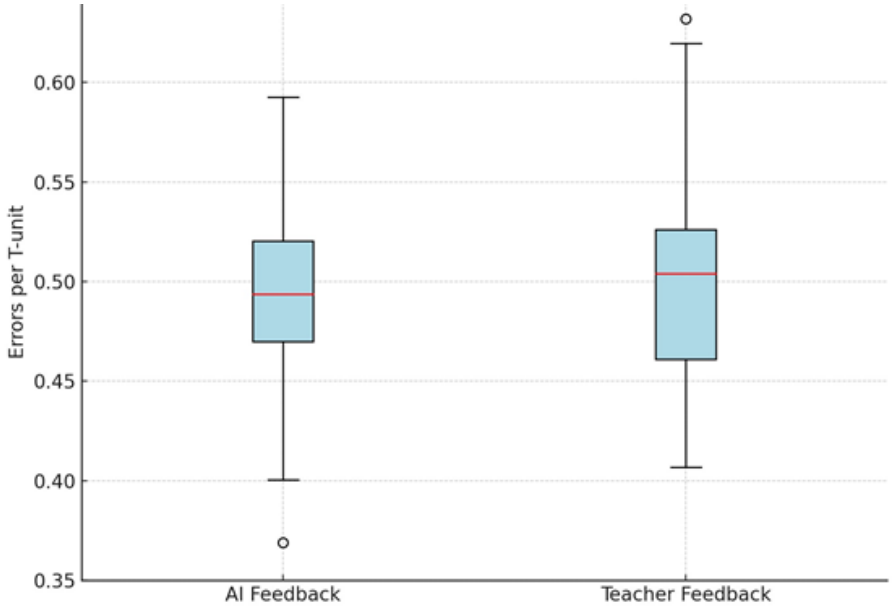


**Figure 2.** *Pre-test scores*

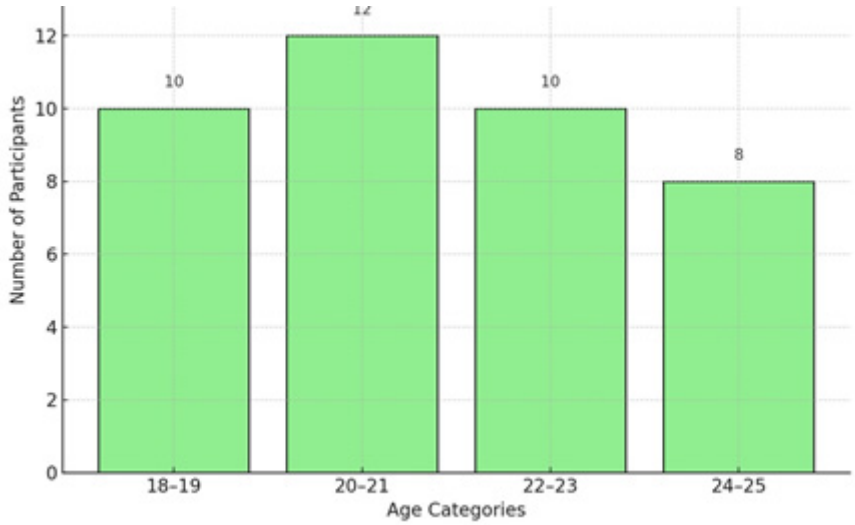


**Figure 3.** *Age distribution*

includes zero, further supporting the equivalence of the groups. Levene's test for equality of variances was non-significant (F = 0.015, p = 0.904), confirming that the assumption of equal variances was met.

To further illustrate the participant profile, a bar chart of age distribution was generated in Figure 3.

This chart confirms that the age distribution was similar between groups, further supporting the comparability of the participant profiles.

### Statistical Analysis of Feedback Efficacy

This section presents the core statistical analysis to evaluate the efficacy of AI-assisted and teacher corrective feedback on grammatical accuracy, addressing the third research question.

### Normality Assessment

The assumption of normality for pre-test and post-test grammatical accuracy scores was tested using the Kolmogorov-Smirnov (K-S) test in SPSS, as parametric tests like ANCOVA require normally distributed data. The results are presented in Table 8.

All p-values are greater than 0.05, indicating that the pre-test and post-test scores for both groups are normally distributed. This satisfies the normality assumption for subsequent parametric tests, ensuring the appropriateness of using ANCOVA and t-tests. The post-test essay topic was: "Some people believe the aim of university education is to help graduates get better jobs. Others believe there are many wider benefits of university education for both individuals and society. Discuss both views and give your own opinion."

### ANCOVA Results

ANCOVA was conducted to compare post-test grammatical accuracy scores between the AI and teacher feedback

**Table 9.** ANCOVA Results for Post-test Grammatical Accuracy Scores

| Source | df | F | p | Partial η² | Adjusted Mean |
|---|---|---|---|---|---|
| Pre-test Score (Covariate) | 1 | 45.672 | <0.001 | 0.559 | -- |
| Group | 1 | 2.333 | 0.036 | 0.066 | -- |
| AI Feedback | -- | -- | -- | -- | 0.4375 |
| Teacher Feedback | -- | -- | -- | -- | 0.4280 |
| Error | 36 | -- | -- | -- | -- |

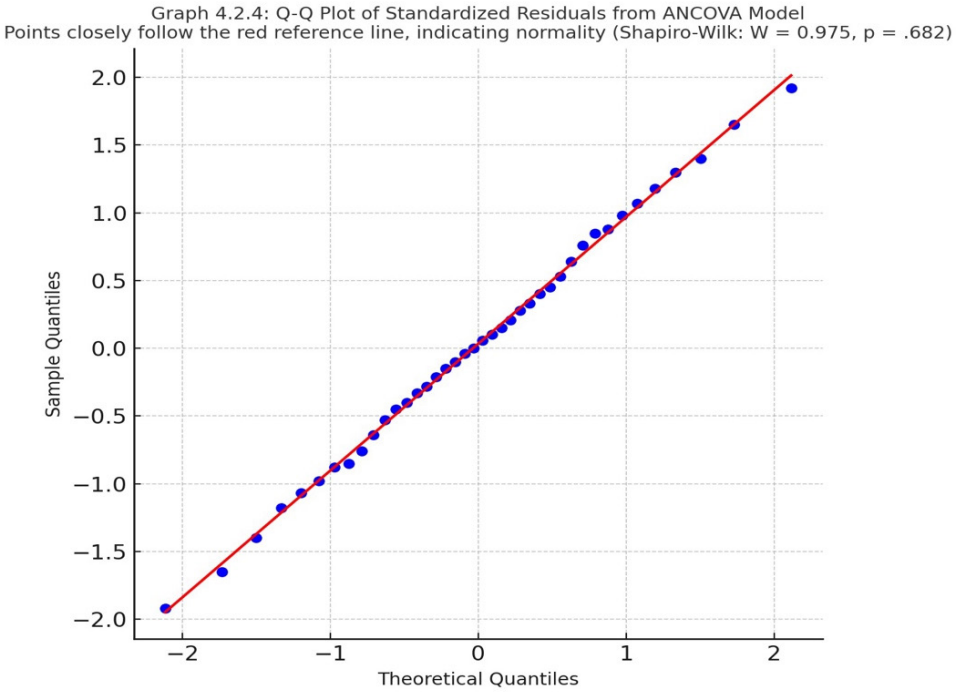


**Figure 4**. Q-Q Plot of Standardized Residuals from ANCOVA Model

groups, with pre-test scores as a covariate to control for baseline performance. The results are presented in Table 9. The ANCOVA revealed a significant group effect ($F(1, 36) = 2.333$, $p = 0.036$, partial $\eta^2 = 0.066$), indicating a statistically significant difference in post-test grammatical accuracy between the AI and teacher feedback groups. The adjusted mean for the teacher feedback group ($M = 0.4280$) was lower than that for the AI feedback group ($M = 0.4375$), suggesting that teacher feedback resulted in slightly better grammatical accuracy (fewer errors per T-unit). The effect size (partial $\eta^2 = 0.066$) indicates a medium effect, suggesting that the difference is practically meaningful but not large. The covariate, pre-test score, was highly significant ($F(1, 36) = 45.672$, $p < 0.001$), confirming that baseline performance significantly influenced post-test scores.

The combination of the Q-Q plot and the Shapiro-Wilk test provides strong evidence that the ANCOVA model is statistically appropriate, reinforcing confidence in the finding that teacher feedback was slightly more effective in improving grammatical accuracy.

*Points closely follow the red reference line, indicating normality (Shapiro-Wilk: W = 0.975, p = .682).* These results confirm that all ANCOVA assumptions were met, supporting the reliability and validity of the findings. The significant group effect ($p = 0.036$) suggests that teacher corrective feedback was slightly more effective than AI-assisted feedback in improving grammatical accuracy among Iranian EFL learners.

**Frequencies of Grammatical Accuracy Improvements**

This section addresses the first and second research questions, examining whether AI-assisted feedback and teacher corrective feedback significantly improved grammatical accuracy from pre-test to post-test. Paired-samples t-tests were conducted to compare pre-test and post-test scores within each group, and descriptive statistics provide insight into the patterns of error reduction.

**Table 10.** Descriptive Statistics of Grammatical Accuracy Scores

| Group | Test | N | Mean | SD | Mean Difference |
|---|---|---|---|---|---|
| AI Feedback | Pre-test | 20 | 0.5000 | 0.0500 | -- |
| AI Feedback | Post-test | 20 | 0.4375 | 0.0450 | 0.0625 |
| Teacher Feedback | Pre-test | 20 | 0.5000 | 0.0485 | -- |
| Teacher Feedback | Post-test | 20 | 0.4280 | 0.0420 | 0.0720 |

**Table 11.** Paired-Samples T-test for AI Feedback Group

| Comparison | t | df | p | Mean Difference | Cohen's d |
|---|---|---|---|---|---|
| Pre-test vs. Post-test | 5.432 | 19 | <0.001 | 0.0625 | 1.30 |

**Table 12.** Paired-Samples T-test for Teacher Feedback Group

| Comparison | t | df | p | Mean Difference | Cohen's d |
|---|---|---|---|---|---|
| Pre-test vs. Post-test | 6.789 | 19 | <0.001 | 0.0720 | 1.58 |

**Table 13.** Pearson Correlation for Pre-test Accuracy Scores

| Rater Pair | Pearson's r | p |
|---|---|---|
| Rater 1 vs. Rater 2 | 0.975 | <0.001 |

**Table 14.** Pearson Correlation for Post-test Accuracy Scores

| Rater Pair | Pearson's r | p |
|---|---|---|
| Rater 1 vs. Rater 2 | 0.980 | <0.001 |

Both groups demonstrated a reduction in errors per T-unit from pre-test to post-test, indicating improvements in grammatical accuracy. The AI feedback group reduced errors by 0.0625 (from 0.5000 to 0.4375), while the teacher feedback group reduced errors by 0.0720 (from 0.5000 to 0.4280). The teacher feedback group exhibited a slightly greater improvement, consistent with the ANCOVA findings.

The t-test was significant ($t(19) = 5.432$, $p < 0.001$), indicating that AI-assisted feedback significantly improved grammatical accuracy. The mean difference of 0.0625 reflects a reduction in errors per T-unit, and the large effect size (Cohen's $d = 1.30$) suggests that the improvement was substantial. This finding confirms that AI feedback, delivered through ChatGPT, was effective in enhancing learners' ability to produce grammatically accurate written texts.

The t-test was significant ($t(19) = 6.789$, $p < 0.001$), confirming that teacher corrective feedback significantly improved grammatical accuracy. The mean difference of 0.0720 indicates a slightly larger reduction in errors compared to the AI group, and the effect size (Cohen's $d = 1.58$) is larger, suggesting a more pronounced improvement. This result highlights the effectiveness of teacher feedback in reducing grammatical errors in written performance.

### Inter-rater Reliability for Accuracy Scores

Two trained raters scored the pre-test and post-test essays. To ensure the reliability of the grammatical accuracy scores, inter-rater reliability was assessed using Pearson's correlation coefficient.

The Pearson correlation coefficient ($r = 0.975$, $p < 0.001$) indicates a very high level of agreement between the two raters for pre-test scores.

The post-test correlation ($r = 0.980$, $p < 0.001$) is similarly high, indicating excellent agreement between raters.

## DISCUSSION

The effectiveness of teacher corrective feedback compared to AI-assisted feedback in improving the grammatical accuracy of Iranian EFL learners' written performance can be attributed to several interrelated factors, with the interactive nature of the classroom setting playing a pivotal role. This approach deepened learners' conceptual understanding, contributing to the significant reduction in errors per T-unit (from 0.5000 to 0.4280, $t(19) = 6.789$, $p < 0.001$, Cohen's $d = 1.58$). The large effect size ($d = 1.58$) indicates a substantial practical impact, suggesting that teacher feedback not only improved accuracy but also fostered meaningful learning gains that could translate to improved performance in high-stakes writing tasks, such as university entrance exams, critical in Iran's EFL context (Rahimi & Mahboob, 2010). In contrast, the AI-assisted feedback group, using ChatGPT, received immediate but text-based suggestions that could be accessed independently. While ChatGPT's feedback was accurate for errors like subject-verb agreement (e.g., "She go" corrected to "She goes"), it lacked interactive scaffolding. The AI group's improvement (from 0.5000 to 0.4375, $t(19) = 5.432$, $p < 0.001$, Cohen's $d = 1.30$) was also substantial, with a large effect size ($d = 1.30$) indicating that AI feedback significantly enhanced accuracy. However, the slightly smaller effect size compared to the teacher group suggests that the absence of real-time dialogue may have limited learners' ability to fully internalize corrections, particularly for complex or context-specific errors. The results are in line with previous studies about the effectiveness of AI on developing written performance (e.g., Attal, et al, 2025; Carla, et al, 2024; Sakai et al, 2023; Tuma et al ,2024; Vaishya et al, 2024). However, this study highlights the effectiveness of teacher feedback over AI-based corrective feedback.

Learner engagement and responsiveness further influenced differential effectiveness. In the teacher feedback group, informal observations noted higher engagement due to personalized explanations and peer discussions. The AI group, while benefiting from ChatGPT's accessibility, occasionally found explanations overly technical (e.g., "Use 'the' for definite nouns"), leading to superficial application of feedback.

Technological limitations of LLMs ( Large Language Models) also contributed to the differential effectiveness. Teachers adapted feedback to learners' proficiency and cultural backgrounds, addressing errors like article omission (common due to Persian's lack of articles) with culturally relevant examples (e.g., "Use 'the' before 'Tehran University' because it's a specific place"). ChatGPT's standardized suggestions (e.g., "Add 'the' before a specific noun") sometimes failed to address nuanced or culturally specific errors, reducing their applicability. The ANCOVA's medium effect size (partial $\eta^2 = 0.066$) suggests that teacher feedback's advantage, while statistically significant, may have a moderate practical impact, indicating potential for hybrid models combining AI efficiency with teacher interaction. This moderate effect size implies that while teachers' context-sensitive feedback yielded slightly better outcomes, AI feedback's scalability could address resource constraints in Iran's overcrowded classrooms, where teacher-to-student ratios often exceed 30:1 (Rahimi & Mahboob, 2010).

The collaborative learning environment in the teacher feedback group, aligned with sociocultural learning theories (Vygotsky, 1978), enhanced effectiveness. Peer discussions and teacher-facilitated interactions created a supportive context where learners reinforced grammatical understanding through collaborative problem-solving. The AI group's solitary interaction with ChatGPT, despite its large effect size ($d = 1.30$), lacked this social component, potentially limiting motivational impact in Iran's collectivist educational culture. The findings support sociocultural learning theories, particularly Vygotsky's (1978) Zone of Proximal Development (ZPD), which posits that learning is most effective when supported by social interaction. Teacher feedback, delivered through dialogue and tailored to individual needs, facilitated deeper rule comprehension within learners' ZPD. In Iran's EFL context, where teacher-centered instruction dominates, the slight superiority of teacher feedback aligns with cultural expectations of teachers as authoritative figures.

The findings inform teacher training and curriculum design in Iranian EFL programs. Training teachers to integrate LLMs ( Large Language Models) like ChatGPT could create hybrid feedback systems, leveraging the large effect sizes of both methods ($d = 1.30, 1.58$) to enhance efficiency without sacrificing interaction. For example, AI could provide initial corrections, followed by teacher-led discussions to clarify complex errors. This approach could scale feedback delivery in Iran's public schools, where technological disparities and budget constraints limit AI adoption (Dashtestani & Hojatpanah, 2020). Future research should explore longitudinal effects to determine if AI feedback's effectiveness ($d = 1.30$) increases with learner familiarity, investigate learner perceptions to enhance engagement, and test hybrid models to optimize outcomes across proficiency levels and error types.

The present findings—showing that writing curricula should adopt a hybrid-feedback model in which automated suggestions handle high-frequency, rule-based errors (verb tense, articles, prepositions), freeing teachers to focus on discourse-level issues and affective support; such a division of labour echoes current CALL advice (Li, Zhu, & Li, 2022) and aligns with task-level feedback principles (Hattie & Timperley, 2007). Second, institutes must provide professional-development workshops on prompt engineering and error-type calibration so that teachers can curate ChatGPT outputs, avoid over-correction, and model metalinguistic explanations—skills that research links to greater uptake (Dashtestani & Hojatpanah, 2020). Third, given Iran's uneven digital infrastructure, administrators should guarantee classroom Wi-Fi (≥ 10 Mbps per 20 devices) and formulate equity guidelines so that learners without personal laptops can still access AI tools during supervised sessions. Fourth, assessment units are advised

to revise rubrics by weighting error categories according to communicative salience (verb tense 3, agreement/prepositions 2, articles/word order 1) and by incorporating AI-generated analytics to inform continuous assessment, thereby promoting criterion-referenced transparency (Knoch, 2009). Finally, policy makers should embed AI ethics clauses—data privacy, academic-integrity checkpoints, and human-in-the-loop review—into national ELT guidelines, ensuring that large-language-model use enhances, rather than replaces, teacher agency. Practically, EFL curricula in Iran should integrate AI-assisted feedback, particularly in urban institutes with technological resources, using tools like ChatGPT for routine corrections to reduce teacher workload (Attal, et al, 2025).

## CONCLUSION AND IMPLICATIONS

This study provides compelling evidence that both AI-assisted feedback (via ChatGPT) and teacher-provided corrective feedback significantly improve grammatical accuracy in Iranian EFL learners' written performance, with teacher feedback demonstrating a slight advantage. The AI feedback group achieved a substantial reduction in errors per T-unit, while the teacher feedback group exhibited a slightly greater improvement. These findings, grounded in a six-week intervention at a Tabriz language institute, contribute to the limited research on AI feedback in Iran's resource-constrained, teacher-centric EFL context. By interpreting these results through sociocultural learning theories (Vygotsky, 1978) and prior studies (e.g., Dashtestani & Hojatpanah, 2020; Li et al., 2022), the study highlights the interactive advantage of teacher feedback and the scalability of AI feedback, offering a nuanced understanding of their roles in enhancing exam-relevant accuracy critical for Iranian learners (Rahimi & Mahboob, 2010). Despite its contributions, the study's small sample size, short intervention duration, and urban focus limit its generalizability, necessitating further research and practical strategies to optimize feedback practices in diverse EFL settings.

The study's implications suggest several avenues for advancing EFL instruction in Iran and similar contexts, particularly through hybrid feedback models that combine AI's efficiency with teacher interaction. To build on these findings, future research should explore longitudinal effects to determine whether AI feedback's effectiveness increases with prolonged exposure, as learners adapt to tools like ChatGPT, potentially narrowing the gap with teacher feedback. Such studies could assess retention of grammatical accuracy and transfer to other writing tasks, addressing the limitation of the six-week intervention (Dashtestani & Hojatpanah, 2020). Additionally, investigating learners' perceptions of AI versus teacher feedback through surveys or interviews could clarify engagement differences, particularly in Iran's teacher-centric culture, where learners value authoritative instruction (Rassouli & Salari, 2018). This could inform strategies to enhance AI's appeal, making it more intuitive for learners accustomed to human interaction. Testing hybrid feedback models, where AI provides initial corrections followed by teacher-led discussions (Attal, et al, 2025; Carla, et al, 2024; Sakai et al, 2023; Tuma et al ,2024; Vaishya et al, 2024), could leverage the large effect sizes of both methods to optimize outcomes in large classes, addressing Iran's high teacher-to-student ratios (Rahimi & Mahboob, 1998). Expanding the focus to include additional grammatical structures, such as relative clauses or modals, would assess feedback efficacy across diverse error types, overcoming the study's narrow scope (Swan, 2005). Replicating the study in rural or public schools, where technological disparities are pronounced, would enhance generalizability beyond urban institutes, validating the findings in diverse contexts (Dashtestani & Hojatpanah, 2020).

Practically, EFL curricula in Iran should integrate AI-assisted feedback, particularly in urban institutes with technological resources, using tools like ChatGPT for routine corrections to reduce teacher workload. This allows instructors to focus on higher-order skills like argumentation, critical for Iran's high-stakes exams (Rahimi & Mahboob, 2010). Teacher training programs should equip instructors with skills to use LLMs ( Large Language Models) effectively, covering prompt design and strategies to scaffold AI corrections through discussions, fostering hybrid systems that leverage the moderate comparative difference to enhance outcomes (Taghizadeh & Yourdshahi, 2020). Implementing hybrid feedback models in classrooms, where AI handles initial corrections and teachers facilitate follow-up interactions, could address large class sizes and resource constraints, combining AI's efficiency with teacher feedback's interactivity (Pishghadam & Zabihi, 1992; Jordan & Mitchell, 2025). Policymakers should invest in technological infrastructure, such as internet access and computer labs, particularly in public schools, to support AI adoption, with offline AI solutions bridging disparities to ensure equitable access (Attal, et al, 2025; Dashtestani & Hojatpanah, 2020). Collaboration with AI developers to tailor tools to Iran's EFL context, incorporating culturally relevant examples and interactive features, could enhance AI feedback's effectiveness, narrowing the gap with teacher feedback (Dashtestani & Hojatpanah, 2020). These strategies would optimize grammatical accuracy outcomes, preparing learners for academic and professional success while addressing practical constraints in Iran's EFL landscape.

Despite its robust findings, the study has several limitations that temper its conclusions. The sample of 40 learners, while sufficient for detecting medium-to-large effects, is small, limiting generalizability to Iran's diverse EFL population. The urban setting in Tabriz, with reliable internet and technological resources, may overestimate AI feedback's effectiveness compared to other cities or private schools with limited infrastructure (Dashtestani & Hojatpanah, 2020). The five-week intervention, though effective, may not capture long-term effects or learner adaptation to AI feedback, necessitating longitudinal studies (Ranalli, 2021). The absence of learner perception

data limits insights into engagement differences, particularly given Iran's cultural preference for teacher authority (Dashtestani & Hojatpanah, 2020). Finally, the researcher's role in coordinating AI feedback and data analysis, despite efforts to minimize bias, could introduce subtle influences, suggesting the need for multiple researchers in future studies ( Lin & Crosthwaite, 2024). By addressing these limitations through expanded research and targeted practical applications, the field can advance toward more effective, context-sensitive feedback practices that enhance EFL learning outcomes globally.

**Authors' contributions**
All authors have contributed equally to prepare the paper.

**Availability of data and materials**
The data that support the findings of this study are available from the corresponding author, upon reasonable request.

**Conflict of interests**
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## REFERENCES

Atai, M. R., & Mazlum, F. (2013). English language-teaching curriculum in Iran: Planning and practice. The Curriculum Journal, 24(3), 389–411. https://doi.org/10.1080/09585176.2012.744326

Attal, A., Shvartz, S., Nakhoul , A.& Daniel Bahir, D.(2025). Chat GPT 4o vs residents: French language evaluation in ophthalmology. AJO International, 2, 34-45. https://doi.org/10.1016/j.ajoint.2025.100104

Carla, M.M., Gambini, G., Baldascino, A. (2024).Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefe's Archive for Clinical and Experimental Ophthalmol, 262(9), 2945–2959. https://doi.org/10.1007/s00417-024-06470-5

Ellis, R. (2009). A typology of written corrective feedback types. ELT Journal, 63(2), 97–107. https://doi.org/10.1093/elt/ccn023

Farhady, H., Sajadi Hezaveh, F., & Hedayati, H. (2010). Reflections on Foreign Language Education in Iran. The Electronic Journal for English as a Second Language, 13(4), 1-18.

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

Hojatpanah, S., & , Dashtestani, R.(2020). Electronic Dictionaries as Language Learning Tools for Iranian Junior High School Students. Computer-Assisted Language Learning Electronic Journal, 21(1). 1-14.

Hyland, K., & Hyland, F. (2019). Feedback in second language writing: Contexts and issues (2nd ed.). Cambridge University Press.

Jordan, M. I. , & Mitchell, T. M.(2025). Machine learning: Trends, perspectives, and prospects. science, 349( 6245), 255-260. https://doi.org/10.1126/science.aaa8415

Knoch, U. (2009). Diagnostic assessment of writing: A review of research and practice. Assessing Writing, 14(2), 112–129. https://doi.org/10.1016/j.asw.2009.04.001

Li, S., Zhu, W., & Li, Y. (2022). The effects of different types of corrective feedback on L2 writing. Language Teaching Research, 26(4), 715–738. https://doi.org/10.1177/1362168820926698

Lin, SH., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. System, 127, 1-17. https://doi.org/10.1016/j.system.2024.103529

Luckin, R., & Holmes, W.(2016). Intelligence Unleashed: An argument for AI in Education. UCL Knowledge

Lab.Lyster, R., & Saito, K. (2010). Oral Feedback in Classroom: SLA. Cambridge University Press. https://doi.org/10.1017/S0272263109990520

Mulyono, H., & Halim, A. (2020). Developmental sequence analysis in Indonesian EFL writing: Implications for instruction. Indonesian Journal of Applied Linguistics, 10(1), 89–100. https://doi.org/10.17509/ijal.v10i1.25012

Pishghadam, R. , Zabihi, R., & Kermanshahi, P.(2012). Educational Language Teaching: A New Movement beyond Reflective/Critical Teaching. Life Science Journal, 9 (1) 1, 892-899.

Rahimi, M., & Mahboob, A. (2010). English as a foreign language in Iran: Challenges and prospects. Language Policy, 9(3), 231–249. https://doi.org/10.1007/s10993-010-9174-6

Ranalli, J. (2021). Automated written corrective feedback: The role of metalinguistic information. Language Learning & Technology, 25(2), 44–61. https://www.lltjournal.org/item/10125-72717

Sakai, D., Maeda, T., Ozaki, A., Kanda, G.N., Kurimoto, Y., Takahashi, M. (2023).Performance of ChatGPT in board examinations for specialists in the Japanese ophthalmology Society. Cureus, 20, 23-34. https://doi.org/10.7759/cureus.49903

Selinker, L. (1972). Interlanguage . IRAL, 10, 209-231. https://doi.org/10.1515/iral.1972.10.1-4.209

Taghizadeh, M., & Yourdshahi, Z. (2020). Technology-enhanced feedback in EFL writing: Teachers' and learners' perspectives. System, 91, 102118. https://doi.org/10.1016/j.system.2020.102118

Tao, B.K., Handzic, A., Hua, N.J., Vosoughi, A.R., Margolin, E.A., Micieli, J.A.(2024). Utility of ChatGPT for automated creation of patient education handouts: an application in Neuro-ophthalmology. Journal of Neuro-ophthalmol, 28, 11-23.

Tuma, N.J., Caterini, J., Liblk, K. (2024). Performance of artificial intelligence on a simulated Canadian urology board exam. Canada Urological Association Journal, 18(10), 22-34. https://doi.org/10.5489/cuaj.8800. 26.

Vaishya, R., Iyengar, K., Patralekh M. (2024).Effectiveness of AI-powered chatbots in responding to orthopaedic postgraduate exam questions—an observational study. International Orthopaedics, 48(8), 1963–1969. https://doi.org/10.1007/s00264-024-06182-9.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Wang, SH., Wang, F.,& Zhu, Z.(2024). Artificial intelligence in education: A systematic literature review. Expert Systems with Applications, 252(4), 124-167. https://doi.org/10.1016/j.eswa.2024.124167

Youn, CH., Salam, A. R., & Rahman, A.A. (2025). AI-Driven Tools in Providing Feedback on Student's Writing. International Journal of Research and Innovation in Social Science, 9(3), 58-67. https://doi.org/10.47772/IJRISS.2025.903SEDU0006