DREAN Teach Me: An Intelligent Tutoring System with NLP and Adaptive Feedback for Elementary School Students

Betty Cotrina
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, San Isidro, Lima, Peru
u202113356@upc.edu.pe

Bryan Perez
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, San Isidro, Lima, Peru
u201723123@upc.edu.pe

Sandra Wong-Durand
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, San Isidro, Lima, Peru
pcsiswon@upc.edu.pe (corresponding author)

Pedro Castaneda
Faculty of Systems Engineering and Electrical Mechanics, Universidad Nacional Toribio Rodriguez de Mendoza, Amazonas, Peru
pedro.castaneda@untrm.edu.pe

Alejandra Onate-Andino
Escuela Superior Politecnica de Chimborazo (ESPOCH), Riobamba, Ecuador
monate@espoch.edu.ec

Received: 25 November 2025 | Revised: 21 January 2026, 19 February 2026, 27 February 2026, and 6 March 2026 | Accepted: 8 March 2026

 DOI: https://doi.org/10.48084/etasr.16487

ABSTRACT

Automated instructional reinforcement is a promising approach to improving learning in elementary education, particularly when personalized and continuous feedback is required. This work presents a web-based intelligent tutoring application leveraging Natural Language Processing (NLP) that delivers automatic feedback based on students' responses to subject-area assessments. The system was applied in Mathematics, Language Arts, and Science & Technology. Effectiveness was evaluated using a quasi-experimental single-group design with pretest and posttest measurements, comparing performance before and after the intervention. The tutoring system adapts reinforcement to baseline performance and operates through graduated levels of support. To estimate the impact, we employed a paired t-test and Cohen's d effect size. The results indicate a mean gain of 0.56 points, $p < 0.001$, and $d = 0.73$ (moderate to large), with the greatest improvement in Language Arts. These findings show that automated reinforcement contributed significantly to improved academic performance, being especially effective for communicative competencies. This underscores the potential of intelligent technologies to transform traditional instruction into a more adaptive and effective process.

Keywords-Intelligent Tutoring System (ITS); Natural Language Processing (NLP); educational reinforcement; adaptive learning; elementary education

I. INTRODUCTION

Low academic performance in elementary education often emerges when traditional instructional approaches fail to adequately differentiate instruction or respond to students' individual learning needs. In contexts characterized by large class sizes and substantial teacher workload, opportunities for personalized support and timely feedback are limited, which constrains the consolidation of foundational skills and allows learning gaps to persist without systematic intervention. Comparable challenges have been reported in higher education, where high student–teacher ratios and delayed instructional feedback have been associated with weaker learning outcomes and reduced academic achievement [1].

This situation is particularly critical in Peru. According to the PISA 2022 results, Peruvian students achieved average scores of 391 in mathematics and 408 in both reading and science, remaining below the Organization for Economic Co-operation and Development (OECD) average across all three domains. Between 2018 and 2022, national performance declined by approximately 9 points in mathematics, whereas reading and science showed only modest improvements of around 8 and 4 points, respectively. Moreover, only 34% of students reached at least Level-2 proficiency in mathematics, compared to 50% in reading and 47% in science, and the proportion of high-performing students (Levels 5 or 6) was negligible. Overall, about 41% of students failed to achieve minimum proficiency across all assessed areas [2]. These results highlight the urgency of implementing effective, scalable instructional support mechanisms from the elementary level onward.

In response to such challenges, educational Artificial Intelligence (AI)—particularly Intelligent Tutoring Systems (ITS) combined with Natural Language Processing (NLP)—has gained prominence as a means of modeling student–system interaction and supporting personalized instruction. Prior research has explored AI-enhanced platforms that automate assessment and feedback processes, such as InteractiveClass, an integrated Microsoft Teams environment designed for collaborative hybrid courses [3]. More broadly, digital learning systems increasingly combine intelligent tutoring, adaptive learning, and predictive analytics to anticipate learning difficulties and recommend appropriate content [4]. Recent advances using Large Language Models (LLMs), including GPT-4–based assistants, further enable scalable feedback generation and multimodal explanations for complex tasks [5]. However, most of these solutions have been developed and validated in higher education contexts, whereas designs explicitly targeting elementary education, supported by in-classroom implementations and longitudinal evaluation, remain relatively scarce.

Existing studies primarily emphasize the potential of LLMs to scale formative assessment and feedback in order to alleviate teachers' time constraints. For example, a GPT-4–based web application evaluated in [5] demonstrated high agreement between automated and expert human feedback in multimodal grading tasks ($r \approx 0.94$). Randomized trials in secondary-school writing contexts have shown that LLM-generated feedback can improve revision quality, motivation, and positive learning-related emotions, although with small-to-moderate effect sizes ($d \approx 0.19$–0.36) [6]. Other work has used process data, such as time on task and edit distance, to reveal how behavioral engagement mediates the effectiveness of automated feedback on performance (Proportion of Mediation (POM) ≈ 0.63 and 0.30) [7]. Studies in English as a Foreign Language (EFL) contexts further report high inter-rater reliability between teacher and ChatGPT-supported feedback ($\kappa \approx 0.85$–0.90), alongside positive learner perceptions [8]. Similarly, the use of ChatGPT for automating grading and feedback in programming assignments has shown strong alignment with instructor evaluations and potential efficiency gains [9]. Collectively, these findings suggest that LLM-based feedback can enhance learning outcomes while improving the scalability of formative assessment.

Beyond feedback generation, several studies have deployed AI-based tutoring and support systems in authentic educational settings to address limitations related to low personalization and delayed instructional response. A ChatGPT-based flipped learning guiding approach evaluated with 81 pre-service teachers significantly outperformed traditional flipped instruction in terms of project quality and learner-related outcomes (partial η^2 up to 0.09) [10]. In another study, an AI learning companion integrated into a Discord platform and combined with Mandala Chart scaffolding produced substantial gains in information-literacy self-efficacy and self-regulated learning among 93 first-year university students (partial $\eta^2 =$ 0.91 and 0.92) [11]. Cross-modal automated evaluation systems for self-directed language learning have also demonstrated low grading error (Mean Absolute Error (MAE) ≈ 0.40 on a 10-point scale) and significant improvements in descriptive performance [12]. Additionally, data-driven ITS approaches trained on large learner datasets have reported high predictive accuracy when modeling academic outcomes using machine learning classifiers, including deep learning architectures [13].

A complementary research stream leverages AI and learning analytics to adapt learning pathways and move beyond one-size-fits-all instruction. Deep-learning-based adaptive learning platforms have reported improvements of approximately 25% in grades, test scores, and engagement across multiple university courses [14]. Other approaches integrate artificial neural networks with fuzzy decision-support systems to recommend tailored interventions based on study strategies and learning profiles [15]. Large-scale STEM interventions using AI kits and computer vision tools have also demonstrated significant gains in conceptual understanding and learner satisfaction in school settings [16]. Similarly, machine-learning-based analyses of multi-year assessment data have been used to identify key evaluation variables and detect inconsistencies or biases in grading practices [17].

Recent review and survey studies further contextualize these developments by synthesizing trends and challenges associated with AI and generative AI in education. Bibliometric analyses covering thousands of publications highlight the growing prominence of ITS, NLP-based language learning, and educational data mining as core research themes

[18, 19]. At the same time, empirical studies report increasing experimentation with generative AI tools among educators, accompanied by concerns regarding academic integrity, assessment validity, and the need for institutional guidance and professional development [20-22].

Within this landscape, this work proposes and evaluates the web application DREAN Teach Me, which integrates ITS and NLP to support learning in fifth- and sixth-grade elementary students. The system follows a three-level instructional flow: (i) an initial assessment consisting of five questions, (ii) a first reinforcement stage based on preloaded explanations retrieved from a database, and (iii) a second and third reinforcement stage in which NLP techniques analyze student errors to generate automated explanations, followed—when necessary—by a conversational tutoring interaction mediated through the ChatGPT Application Programming Interface (API) with controlled token usage. This structure aims to personalize instruction while providing continuous monitoring through visual performance indicators. The system's impact is assessed using a quasi-experimental single-group pretest–posttest design that compares student performance before and after the intervention.

II. SYSTEM DESIGN

The main contribution of this work is the design and implementation of a web-based educational application that integrates an ITS with NLP to support adaptive reinforcement in elementary education. Figure 1 presents the logical architecture of DREAN Teach Me, which has been designed to support multiple user roles (students, teachers, and administrators) and to operate on standard web browsers (Chrome, Edge, Firefox, and Opera) across desktop, tablet, and mobile devices without requiring local installation.

User access begins with a credential-based authentication process that ensures role-specific access to system functionalities. Once authenticated, all interactions are handled within the application layer, which is implemented as a Single-Page Application (SPA) using Angular for the frontend and PHP for the backend.

The frontend layer organizes the learning process into clearly defined functional modules. The Initial Assessment Module administers a subject-specific diagnostic evaluation consisting of five timed questions in Mathematics, Language Arts, or Science & Technology. Upon completion, the Continuous Assessment Module delivers a second set of five questions, interleaved with adaptive reinforcement. Instructional support is structured into three sequential reinforcement modules: First Reinforcement, which displays a predefined explanation retrieved from the database; Second Reinforcement, which applies ITS logic combined with NLP to generate an automated explanation based on the student's error; and Third Reinforcement, which enables a free-form conversational interaction through an NLP-based chat interface when additional clarification is required. Complementary frontend modules include Content Management, allowing teachers to upload and manage assessment items; User Management, which controls permissions and roles; and a Progress Visualization Module, which presents performance indicators and learning trajectories to instructors.

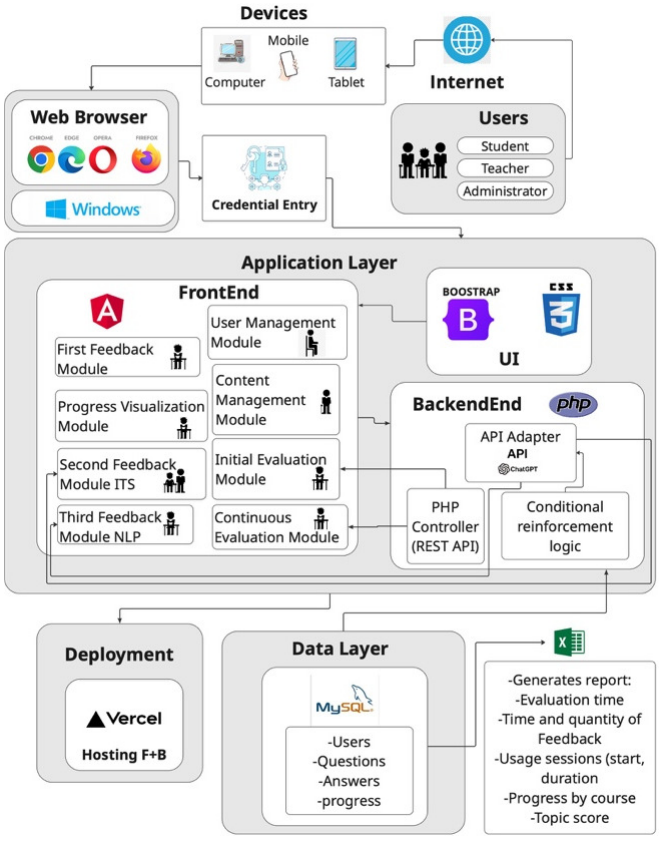

Fig. 1. Logical architecture of the web application based on NLP and ITS.

Backend processing is managed by a PHP controller exposed through a RESTful API. This controller orchestrates assessment logic, reinforcement activation, and data persistence. When second- or third-level reinforcement is triggered, the backend connects to the ChatGPT API to generate context-aware pedagogical explanations, while enforcing predefined constraints on prompt structure and response length to ensure educational relevance and consistency.

The technology and infrastructure layer supports scalability and accessibility. The frontend is deployed on Vercel, leveraging cloud-based hosting for responsive delivery, whereas the backend integrates external services such as the OpenAI API. Persistent data storage is handled by a MySQL database that maintains records of users, assessment items, responses, and learning progress. For instructional monitoring and analysis, the system automatically generates Excel reports summarizing evaluation time, frequency and type of reinforcements, session usage, progress by subject, and topic-level scores.

With this architectural design, the web application provides a modular and extensible tutoring environment in which assessment, adaptive feedback, and monitoring are tightly integrated. By explicitly structuring reinforcement into

progressive levels and coupling ITS decision logic with NLP-based explanation generation, the system moves beyond generic web-based learning platforms toward a controlled and pedagogically oriented adaptive tutoring framework.

III. METHODOLOGY

A. Dataset Collection

The dataset was collected on June 19, 2025, from 102 elementary students (51 fifth-grade and 51 sixth-grade) at Divino Jesús Private School in Comas, Lima, Peru. Data collection took place during regular classroom activities across three subject areas: Mathematics, Communication, and Science & Technology. The instructional bank consisted of 15 multiple-choice items per subject (45 items total), and each item included four answer options with a single correct answer.

For each student, the system administered five items in the Initial Assessment (pretest) and five non-repeated items in the Continuous Assessment (posttest). Item selection was configured to prevent repetition between both assessments for the same student. In total, each student answered 10 items, resulting in 1,020 recorded responses across the full sample.

Collected data included selected answer, binary correctness, response time (in seconds), subject area, grade level, instructional module, and reinforcement phase. The average session duration ranged between 5 and 8 min per student.

In addition to collected interaction data, the system generated derived data during tutoring, including reinforcement level activation (R1, R2, R3), automated explanatory text, conversational logs, and aggregated performance indicators. Approximately 15% of incorrect responses triggered the third reinforcement level (R3). All data were stored in MySQL, pseudonymized using alphanumeric identifiers, and exported to Excel for statistical analysis.

B. Data Preprocessing

The data architecture of DREAN Teach Me is illustrated in Figure 2 and describes the complete information flow from content ingestion to instructional monitoring. The process begins with administrators uploading Excel files containing question banks, answer options, correct responses, and reinforcement texts, as well as managing static educational assets such as images served through signed URLs. These inputs are stored in a structured MySQL database and subsequently consumed by the Assessment, Intelligent Tutoring, and Monitoring modules, ensuring traceability and consistency for administrators, teachers, and students.

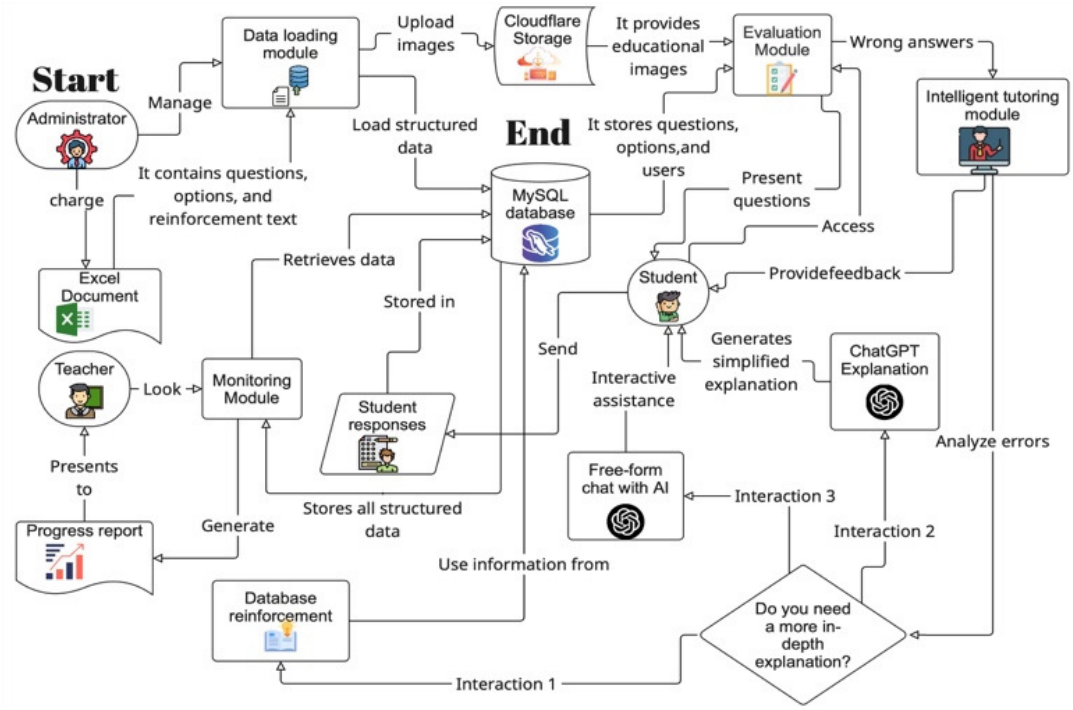

Fig. 2. Data architecture of DREAN Teach Me.

1) Data Ingestion and Validation

During data ingestion, the 45-item bank was validated to verify required fields, data types, value ranges, identifier uniqueness, and question–option integrity (four options per item, one correct answer). Formatting inconsistencies were detected in fewer than 5% of uploaded records and corrected prior to publication.

As part of preprocessing, textual content and pedagogical labels (competency, difficulty level, and cognitive dimension) were normalized using controlled vocabularies. Spacing, capitalization, and accent inconsistencies were corrected to ensure uniformity. Images were handled as external Content

Delivery Network (CDN) references with availability and integrity checks, reducing database load and preventing broken links. Validation reports and change logs were generated to ensure traceability.

2) Feature Extraction

For each student attempt, nine structured features were extracted: (1) grade level, (2) subject area, (3) item identifier, (4) selected option, (5) binary correctness, (6) response time in seconds, (7) reinforcement level activated (R1–R3), (8) session identifier, and (9) timestamp. These variables enabled tracking of learning trajectories and reinforcement dynamics across interactions.

3) Data Transformation

The relational data model defines canonical keys (id_item, id_anon, timestamp) and normalizes tables for items, options, responses, interventions, and users. Analytical views are created to join student attempts with pedagogical metadata. Categorical variables are encoded using ordinal or one-hot representations as appropriate, and aggregated indicators are generated by session, week, and subject area.

For effectiveness analysis, matched pretest–posttest pairs are assembled for each student and subject. The resulting datasets are exported to Excel (CSV/XLSX) and serve as inputs for the statistical analyses described in the following sections.

C. System Development

This subsection focuses on the tutoring framework that orchestrates the student's end-to-end instructional experience rather than on the training of predictive classifiers. The application is implemented as a single-page web system, with an Angular-based frontend and a PHP backend exposing a RESTful API connected to a MySQL database.

The instructional flow begins with the Initial Assessment, which consists of five timed items, followed by a Continuous Assessment phase that interleaves evaluation with adaptive reinforcement. A tutoring module dynamically selects the appropriate intervention level based on student performance. All interactions are logged, including pseudonymized user identifiers, item IDs, timestamps, correctness, and activated reinforcement level, enabling subsequent analysis and reporting. The system is deployed on the web, with the frontend hosted on Vercel, and integrates the ChatGPT API exclusively for generating contextualized pedagogical explanations.

The adaptive pedagogical framework is structured around three reinforcement levels triggered by student responses. The first level provides a baseline explanation retrieved from the database. The second level generates automated feedback through ITS and NLP mechanisms without additional student input: the system analyzes the error made and, using pedagogical labels (competency, difficulty, and cognitive dimension), composes a prompt that includes the original problem statement, associated visual elements, and the selected answer. This prompt is sent to the ChatGPT API to generate an explanation aligned with the intended learning objective (Figure 3).

Fig. 3. Screen showing the second feedback with the NLP and ITS in the application.

If uncertainty persists, the third reinforcement level activates a free-form conversational interface in which the student can express questions in natural language (e.g., "Explain the solution in more detail…"). The system responds with a personalized explanation aimed at fostering self-directed learning and conceptual clarification (Figure 4). In parallel, item banks are ingested and validated from Excel files, images are delivered via CDN links, and teachers access subject-level progress visualizations through a monitoring dashboard.

Fig. 4. Screen showing the third feedback from a free chat with NLP.

D. System Configuration and Quasi-Experimental Validation

This study documents both the configuration of the tutoring system and its in-class validation using a quasi-experimental single-group pretest–posttest design. System configuration includes prompting templates, activation rules governing the transition between reinforcement levels (R1, R2, and R3), and API parameters designed to produce concise, pedagogically focused responses.

The application records pretest data during the Initial Assessment and posttest data during the Continuous Assessment after the reinforcement cycle. Pseudonymized records—comprising student ID, item identifier, response time, correctness, and intervention type—are stored in MySQL and exported to Excel for analysis.

The primary outcome metric is the individual post–pre score difference (Δ), aggregated at the subject and group levels.

Statistical inference relies on paired-sample t-tests to evaluate whether the mean Δ differs significantly from zero at a significance level of $\alpha = 0.05$, accompanied by confidence intervals where applicable.

Effect size is estimated using Cohen's d for paired samples to quantify the magnitude of observed changes independently of sample size. Operational indicators, such as response time and frequency of reinforcement activation, are examined solely to contextualize system usage rather than to infer learning effects.

E. Performance Evaluation Metrics

To evaluate learning gains associated with automated reinforcement, a pretest–posttest evaluation strategy was adopted in which validation metrics are based on direct within-student comparisons. This approach enables the assessment of whether the pedagogical intervention was associated with measurable performance improvement without relying on classification accuracy or prediction metrics.

The central evaluation metric is the per-student score difference (Δ) between posttest and pretest results, with the mean Δ summarizing overall improvement. Statistical significance is assessed using a paired-samples Student's t-test with a two-tailed contrast and a significance threshold of $\alpha = 0.05$. Effect magnitude is quantified using Cohen's d for paired data, facilitating interpretation independently of sample size. When the assumption of normality for Δ is not plausible, the Wilcoxon signed-rank test is employed as a nonparametric alternative (Table I).

As a complementary analysis, Δ values are aggregated by subject area (Mathematics, Language Arts, and Science & Technology), and system usage indicators—such as response time, block-level accuracy, and activation frequency of R2 and R3—are monitored to contextualize the intervention. However, learning inferences are derived exclusively from Δ, the paired t-test, and Cohen's d, consistent with the in-class quasi-experimental design and the objective of isolating the pedagogical effect of the R1→R2→R3 tutoring sequence.

TABLE I. PRE-POST DESIGN PERFORMANCE EVALUATION METRICS AND NOTATION USED

Metric	Description	Formula	Comment
Average (pre), average (post)	Average score at each measurement	—	Baseline comparison (5 items)
Δ (score difference)	Individual gain and average gain	$\Delta_i = post_i - pre_i$; $\bar{\Delta} = \frac{1}{n}\sum \Delta_i$	Primary endpoint of the study. Δ_i: individual difference; $\bar{\Delta}$: average of the differences.
Paired t	Tests whether mean $\bar{\Delta}$ differs from 0	$t = \frac{\bar{\Delta}}{s_\Delta/\sqrt{n}}$	Bilateral test, $\alpha = 0.05$; report p-value and 95% CI. $\bar{\Delta}$: average of the differences; s_Δ: standard deviation of differences; n: number of students.
Cohen's d (paired)	Magnitude of change independent of n	$d = \frac{\bar{\Delta}}{s_\Delta}$	Interpretation: small ≈ 0.2, medium ≈ 0.5, large ≈ 0.8. $\bar{\Delta}$: average of the differences. s_Δ: standard deviation of differences.
95% CI ($\bar{\Delta}$)	Accuracy of the average change estimate	$\bar{\Delta} \pm t_{(1-\frac{\alpha}{2},n-1)} \frac{s_\Delta}{\sqrt{n}}$	Include along with the p-value.
%Δ (percentage improvement)	Relative change from the initial average	$\%\Delta = \frac{\bar{X}_{post} - \bar{X}_{pre}}{\bar{X}_{pre}} \times 100$	Descriptive. Useful for communication; sensitive to low baselines. Not used to infer learning on its own. %Δ: percentage improvement relative to $\bar{X}_{pre}$.
Wilcoxon	Non-parametric alternative for Δ	—	Use when Δ normality is not plausible.

CI = confidence interval.

TABLE II. RESULTS OF INDIVIDUAL SCORES IN THE PRE-TEST AND POST-TEST BY SUBJECT AREA AND GRADE

Grade	Subject	Pre-test total	Post-test total
5th	Mathematics	187	215
5th	Communication	169	211
5th	Science & Technology	194	219
6th	Mathematics	203	215
6th	Communication	186	215
6th	Science & Technology	176	211

IV. RESULTS

To evaluate the impact of the DREAN Teach Me web application on elementary students' academic performance, a quasi-experimental single-group pretest–posttest design was applied. Scores were obtained directly from the system through the Initial Assessment Module (pretest) and the Continuous Assessment Module (posttest), the latter administered after completion of the three reinforcement phases (R1: preloaded explanation; R2: automated explanation via NLP; R3: AI-based conversational support).

The total sample comprised 102 elementary students, with consolidated results reported for fifth and sixth grades (51 students per grade). Individual scores were organized by curricular area—Mathematics, Communication, and Science & Technology—and aggregated by grade and subject (Table II). Across all grade–subject combinations, no decreases in performance were observed after the reinforcement sequence; outcomes showed either stability or improvement.

To assess learning gains associated with automated reinforcement, a pretest–posttest comparison strategy was adopted, consistent with the evaluation framework described in Section III. Performance validation relied on within-student comparisons of scores before and after the intervention. Three complementary metrics were used: (i) the score difference (Δ), computed for each student and averaged to capture overall improvement; (ii) a paired-sample Student's t-test (two-tailed, $p < 0.05$) to evaluate whether the observed mean change differed from zero; and (iii) Cohen's d for paired samples to estimate the magnitude of the change independently of sample size. In the aggregated analysis across subjects, the mean score difference was 0.56, the paired t-test yielded $p < 0.001$, and the effect size was $d = 0.732$, which can be interpreted as moderate-to-high according to conventional thresholds (Table III).

TABLE III. RESULTS OF GENERAL PRE-TEST AND POST-TEST PAIRED SAMPLE STATISTICS

Metric	Pre-test (n = 306)	Post-test (n = 306)	Score difference (Δ)
Average	3.64	4.20	0.56
Standard deviation	1.102	0.510	—
Correlation (pre/post)	—	0.834	—
p-value (two-tailed)	—	< 0.001	—
Cohen's d	—	0.732	—

To further examine whether the effect of adaptive reinforcement was consistent across curricular areas, results were disaggregated by subject. For Mathematics, Communication, and Science & Technology, paired-sample t-tests were conducted to compute the pretest–posttest score difference (Δ), associated p-values, and Cohen's d. All three subjects exhibited statistically significant improvements, with Communication showing the largest average gain (highest Δ) and moderate-to-high effect sizes across subjects (Table IV). These results indicate that posttest averages exceeded pretest scores in all areas.

TABLE IV. COMPARISON OF PRE-TEST AND POST-TEST SCORES BY SUBJECT

Subject	Mathematics	Communication	Science & Technology
Pre-test	3.82	3.48	3.63
Post-test	4.22	4.18	4.22
Improvement (Δ)	+0.40	+0.70	+0.59
p-value	< 0.001	< 0.001	< 0.001
Cohen's d	0.677	0.742	0.749
N	102	102	102
Significance	Yes	Yes	Yes

N = sample size (number of students).

Figure 5 provides a visual comparison of pretest and posttest mean scores by subject. While overall gains are evident, a ceiling effect was observed in Mathematics: a subset of students achieved the maximum score (5) already in the pretest and remained at this level in the posttest. This effect limits the observable increase for these students and may lead to conservative estimates of Δ and Cohen's d in this subject. No comparable ceiling effects were detected in Communication or Science & Technology.

V. DISCUSSION

The results indicate a consistent improvement between pretest and posttest scores, with an average gain of 0.56 points across subjects. This change was statistically significant (paired-sample t-test, $p < 0.001$) and accompanied by a moderate-to-high effect size (Cohen's $d = 0.732$), suggesting that the observed improvement is not only statistically detectable but also educationally meaningful within the context of the intervention. All curricular areas exhibited significant gains ($p < 0.01$), with effect sizes exceeding 0.6. Communication showed the largest improvement ($\Delta = +0.70$; $d = 0.74$), followed by Science & Technology ($\Delta = +0.59$; $d = 0.74$) and Mathematics ($\Delta = +0.40$; $d = 0.67$). Figure 5 visually confirms that posttest mean scores exceeded pretest means across all subjects.

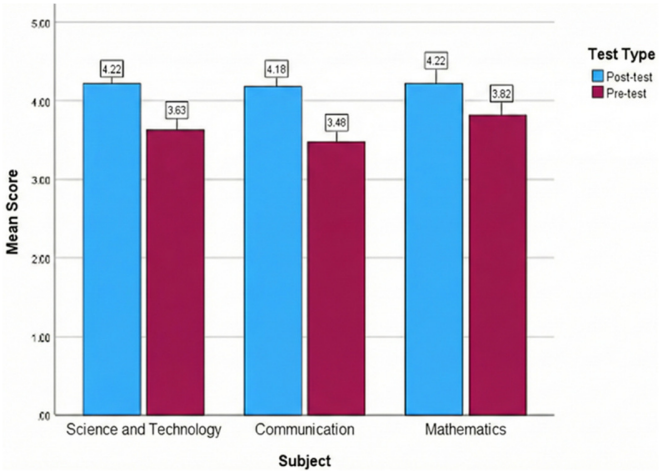

Fig. 5. Increase in average scores after the reinforcement, by course.

For example, in a Mathematics item involving fraction addition with unlike denominators, when students selected an option that added numerators without first obtaining a common denominator, the R2 explanation explicitly guided them to align denominators and recompute the result step by step. If uncertainty persisted, R3 enabled students to request additional clarification in natural language, reinforcing conceptual understanding beyond simple answer correction.

When positioned relative to prior work, the magnitude of the effect observed in this study ($d \approx 0.73$) compares favorably with AI-generated feedback interventions in secondary education, where gains are typically smaller (e.g., $d \approx 0.19$ in writing revision tasks, with motivational effects around $d \approx 0.34$–0.36) [6]. The results are comparable to those reported in ChatGPT-based flipped-classroom approaches with medium-sized improvements (partial η^2 up to 0.09) [10], and align with performance gains observed in higher education AI deployments [3] and adaptive deep-learning platforms reporting improvements of approximately 25% [14].

Unlike many prior studies conducted in higher education or controlled environments [3, 5, 11], the present system was implemented in authentic elementary classrooms using a fixed 45-item question bank (15 per subject), short sessions (5–8 min), and a three-level reinforcement mechanism (R1–R3). This real-world deployment demonstrates practical feasibility in regular school settings rather than experimental laboratory conditions.

From a system-design perspective, the findings align with intelligent tutoring literature emphasizing adaptive, error-contingent feedback. Classical dialogue-based ITS such as AutoTutor [23] demonstrated that conversational scaffolding supports learning through guided reasoning. The present approach extends this principle by integrating controlled NLP-based explanations within a structured ITS framework, enabling scalable feedback while preserving pedagogical alignment for elementary learners.

The stronger effect observed in Communication aligns with prior evidence that language-based tasks benefit from immediate explanatory feedback [12]. The smaller gain in

Mathematics is partly attributable to a ceiling effect, as some students achieved the maximum score in the pretest, limiting measurable improvement. This highlights the need for progressive difficulty scaling in future iterations.

Overall, these findings provide empirical evidence that a structured ITS combined with controlled NLP-based reinforcement can generate measurable academic gains in elementary education within short-duration, real-world classroom sessions.

VI. LIMITATIONS AND FUTURE SCOPE

This study presents several limitations that should be considered when interpreting the results. First, the quasi-experimental pretest–posttest design did not include a control group, which limits the ability to fully isolate the effects of the intervention from potential confounding factors such as maturation, practice effects, or external influences. Although statistical significance and effect sizes provide evidence of improvement, causal claims should therefore be interpreted with caution.

Second, the use of a 0–5 scoring scale and a short test battery of five items per assessment may have introduced ceiling effects, particularly in Mathematics, and reduced sensitivity to detect smaller learning gains. Third, the sample was limited to two grade levels (fifth and sixth grade) and three subject areas, which constrains the generalizability of the findings to other educational levels, subjects, or institutional contexts. Additionally, the evaluation was conducted immediately after the intervention, without a delayed posttest to assess knowledge retention over time. Finally, the system relies on an external NLP API, and aspects such as equity, long-term cost, and robustness to prompt variability warrant further investigation.

Future work will address these limitations along two main directions. First, a more rigorous experimental design incorporating both control and experimental groups will be implemented to strengthen internal validity and enable a clearer estimation of causal effects. Second, the system will be extended with a progressive difficulty mechanism (basic–intermediate–advanced) that activates once students achieve high accuracy levels, thereby reducing ceiling effects and supporting sustained adaptive learning. Additional avenues include longitudinal evaluations, equity-focused analyses, and the application of multilevel models to account for the hierarchical structure of educational data.

VII. CONCLUSION

This paper presented DREAN Teach Me, a web-based educational application that integrates an Intelligent Tutoring System (ITS) with Natural Language Processing (NLP) to provide adaptive reinforcement for elementary school students. The system combines structured instructional flow with automated explanations and optional conversational support, enabling scalable and personalized feedback within regular classroom settings.

Using a quasi-experimental single-group pretest–posttest design, the evaluation showed statistically significant improvements in student performance across Mathematics, Communication, and Science & Technology. The observed learning gains were accompanied by moderate-to-high effect sizes, indicating educational relevance beyond statistical significance. The results further suggest that structuring reinforcement into progressive levels—ranging from predefined explanations to NLP-based automated feedback and conversational tutoring—can support learning while maintaining pedagogical control appropriate for elementary education.

Beyond performance outcomes, this work contributes empirical evidence from authentic school contexts, addressing a gap in the literature where Artificial Intelligence (AI)-driven tutoring systems have been predominantly evaluated in higher education or controlled environments. The findings indicate that integrating ITS principles with NLP-based feedback is feasible in elementary classrooms and can enhance formative assessment without imposing additional workload on teachers.

Nevertheless, the conclusions are bounded by the study's quasi-experimental design and the absence of a control group. As discussed, future research should adopt more rigorous experimental designs, extend the evaluation to additional grade levels and subjects, and incorporate progressive difficulty mechanisms to mitigate ceiling effects and assess long-term retention. Overall, DREAN Teach Me represents a step toward practical, scalable, and pedagogically grounded AI-supported tutoring solutions for elementary education.

DECLARATION OF COMPETING 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.

ACKNOWLEDGMENT

The authors acknowledge the Dirección de Investigación of the Universidad Peruana de Ciencias Aplicadas (UPC) for supporting this work through the UPC-EXPOST-2026-1 incentive.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] H. B. Essel, D. Vlachopoulos, A. Tachie-Menson, E. E. Johnson, and P. K. Baah, "The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education," *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, Nov. 2022, Art. no. 57, https://doi.org/10.1186/s41239-022-00362-6.

[2] OECD, *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. Paris, France: OECD Publishing, 2023, https://doi.org/10.1787/53f23881-en.

[3] W. E. Villegas-Ch, J. Govea, R. Gutierrez, and A. Mera-Navarrete, "Improving Interaction and Assessment in Hybrid Educational Environments: An Integrated Approach in Microsoft Teams With the Use of AI Techniques," *IEEE Access*, vol. 12, pp. 93723–93738, 2024, https://doi.org/10.1109/ACCESS.2024.3424397.

[4] K. S. Suryanarayana, V. S. P. Kandi, G. Pavani, A. S. Rao, S. Rout, and T. Siva Rama Krishna, "Artificial Intelligence Enhanced Digital Learning for the Sustainability of Education Management System," *The*

Journal of High Technology Management Research, vol. 35, no. 2, Nov. 2024, Art. no. 100495, https://doi.org/10.1016/j.hitech.2024.100495.

[5] L. Jürgensmeier and B. Skiera, "Generative AI for scalable feedback to multimodal exercises," *International Journal of Research in Marketing*, vol. 41, no. 3, pp. 468–488, Sept. 2024, https://doi.org/10.1016/j.ijresmar.2024.05.005.

[6] J. Meyer *et al.*, "Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions," *Computers and Education: Artificial Intelligence*, vol. 6, June 2024, Art. no. 100199, https://doi.org/10.1016/j.caeai.2023.100199.

[7] R. Schiller, J. Fleckenstein, U. Mertens, A. Horbach, and J. Meyer, "Understanding the effectiveness of automated feedback: Using process data to uncover the role of behavioral engagement," *Computers & Education*, vol. 223, Dec. 2024, Art. no. 105163, https://doi.org/10.1016/j.compedu.2024.105163.

[8] J. Han and M. Li, "Exploring ChatGPT-supported teacher feedback in the EFL context," *System*, vol. 126, Nov. 2024, Art. no. 103502, https://doi.org/10.1016/j.system.2024.103502.

[9] M. Jukiewicz, "The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process," *Thinking Skills and Creativity*, vol. 52, June 2024, Art. no. 101522, https://doi.org/10.1016/j.tsc.2024.101522.

[10] H. Li, "Effects of a ChatGPT-based flipped learning guiding approach on learners' courseware project performances and perceptions," *Australasian Journal of Educational Technology*, vol. 39, no. 5, pp. 40–58, Dec. 2023, https://doi.org/10.14742/ajet.8923.

[11] Y.-H. Hu, C.-L. Hsieh, and E. S. N. Salac, "Advancing freshman skills in information literacy and self-regulation: The role of AI learning companions and Mandala Chart in academic libraries," *The Journal of Academic Librarianship*, vol. 50, no. 3, May 2024, Art. no. 102885, https://doi.org/10.1016/j.acalib.2024.102885.

[12] R. Zhao, Y. Zhuang, Z. Xie, and P. L. H. Yu, "Facilitating self-directed language learning in real-life scene description tasks with automated evaluation," *Computers & Education*, vol. 219, Oct. 2024, Art. no. 105106, https://doi.org/10.1016/j.compedu.2024.105106.

[13] K. A. Aldriwish, "Empowering Learning through Intelligent Data-Driven Systems," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12844–12849, Feb. 2024, https://doi.org/10.48084/etasr.6675.

[14] F. Naseer, M. N. Khan, M. Tahir, A. Addas, and S. M. H. Aejaz, "Integrating deep learning techniques for personalized learning pathways in higher education," *Heliyon*, vol. 10, no. 11, June 2024, Art. no. e32628, https://doi.org/10.1016/j.heliyon.2024.e32628.

[15] A. Bressane *et al.*, "Understanding the role of study strategies and learning disabilities on student academic performance to enhance educational approaches: A proposal using artificial intelligence," *Computers and Education: Artificial Intelligence*, vol. 6, June 2024, Art. no. 100196, https://doi.org/10.1016/j.caeai.2023.100196.

[16] M. Lohakan and C. Seetao, "Large-scale experiment in STEM education for high school students using artificial intelligence kit based on computer vision and Python," *Heliyon*, vol. 10, no. 10, May 2024, Art. no. e31366, https://doi.org/10.1016/j.heliyon.2024.e31366.

[17] J. Divasón, F. J. Martínez-de-Pisón, A. Romero, and E. Sáenz-de-Cabezón, "Artificial Intelligence Models for Assessing the Evaluation Process of Complex Student Projects," *IEEE Transactions on Learning Technologies*, vol. 16, no. 5, pp. 694–707, Oct. 2023, https://doi.org/10.1109/TLT.2023.3246589.

[18] X. Chen, D. Zou, H. Xie, G. Cheng, and C. Liu, "Two Decades of Artificial Intelligence in Education: Contributors, Collaborations, Research Topics, Challenges, and Future Directions," *Educational Technology & Society*, vol. 25, no. 1, pp. 28–47, Jan. 2022, https://doi.org/10.30191/ETS.202201_25(1).0003.

[19] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du, "Artificial intelligence in education: A systematic literature review," *Expert Systems with Applications*, vol. 252, Oct. 2024, Art. no. 124167, https://doi.org/10.1016/j.eswa.2024.124167.

[20] D. Lee *et al.*, "The impact of generative AI on higher education learning and teaching: A study of educators' perspectives," *Computers and Education: Artificial Intelligence*, vol. 6, June 2024, Art. no. 100221, https://doi.org/10.1016/j.caeai.2024.100221.

[21] T. K. F. Chiu, "The impact of Generative AI (GenAI) on practices, policies and research direction in education: a case of ChatGPT and Midjourney," *Interactive Learning Environments*, vol. 32, no. 10, pp. 6187–6203, Nov. 2024, https://doi.org/10.1080/10494820.2023.2253861.

[22] R. F. Kizilcec *et al.*, "Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States," *Computers and Education: Artificial Intelligence*, vol. 7, Dec. 2024, Art. no. 100269, https://doi.org/10.1016/j.caeai.2024.100269.

[23] A. C. Graesser, K. VanLehn, C. P. Rose, P. W. Jordan, and D. Harter, "Intelligent Tutoring Systems with Conversational Dialogue," *AI Magazine*, vol. 22, no. 4, pp. 39–39, Dec. 2001, https://doi.org/10.1609/aimag.v22i4.1591.