AI-Driven Education: Integrating Machine Learning and NLP to Transform Child Learning Systems

Mst Masuma Akter Semi[1], Md Borhan Uddin[2], Sharmin Sultana[3], Motmainna Tamanna[4], Azim Uddin[5], Khandakar Rabbi Ahmed[6]*

MA in TESOL, Westcliff University, Irvine, CA, USA[1]
Department of Business Administration and Management, International American University, Los Angeles, USA[2, 3]
Bachelor of Business Management, University of Portsmouth, United Kingdom[4]
Bachelor of Business Administration, Southern California State University, Los Angeles, USA[5]
Miyan Research Institute (MRI), International University of Business Agriculture and Technology, Dhaka, Bangladesh[6]

***Abstract*—An Artificial Intelligence-driven child learning system with a Machine Learning and Natural Language Processing-based approach to dynamically personalize educational experiences for children is proposed in this study. Using a Sentence-BERT model to encode student queries for the computation of semantic similarity and knowledge domains to be retrieved. A T5-based transformer model writes verbose, personalized feedback, and a Gradient Boosting Machine classifier predicts the appropriate learning outcomes. The content difficulty and personalization of educational trajectories across content are set by an integrated adaptive learning engine that monitors and adjusts for student performance. On the General Knowledge QA dataset, classification accuracy reaches 85.2%, and the ROC-AUC score is 0.912, which has been proven to be reliable in real-world cases. It also produces positive effects regarding the understanding and preference for learners of adaptive systems, as observed in user studies. AI technologies have exciting potential to deliver scalable, personalized education for young learners, as demonstrated in this work.**

Keywords*—Artificial intelligence; machine learning; natural language processing; adaptive learning systems; sentencebert; gradient boosting machine; personalized feedback

I. Introduction

The adoption of Artificial Intelligence (AI) [1], Machine Learning (ML) [2], and Natural Language Processing (NLP) [3] technologies is almost revolutionizing the current landscape of education [4]. Traditional approaches to learning system design aspire to a single solution for teaching and learning that is not tailored to the individual children's unique needs, learning speeds, or knowledge gaps [5]. Fortunately, by taking advantage of these challenges, AI-based education offers the opportunity to personalize learning, assess student capabilities in adaptive ways, and provide personalized feedback to enhance both engagement and evaluation [6].

Now, various advances in deep learning (DL) and NLP have enabled systems to process natural language and learn complex behaviors through real-time feedback [7]. In proposing knowledge assessment and content tailoring based on automation, techniques such as sentence embeddings, semantic similarity measures, and gradient-boosted decision trees have demonstrated either great promise or practicality. Moreover, transformer-based models [8], with T5 (Text-to-Text Transfer Transformer) being one of them, have led to an increase in feedback generation through their high-end, contextual, and human-like responses that enhance the entire learning experience.

In this study, we proposed an AI-driven child learning system that uses ML and NLP to adapt to a student's learning progression dynamically. The system first encodes student queries into a sentence embedding model (SBERT) and computes semantic similarity to identify the most related knowledge domain using the General Knowledge Q&A dataset.

The appropriate learning outcomes are predicted by a Gradient Boosting Machine (GBM) classifier, and a T5-based feedback engine generates the detailed, adaptive feedback. The adaptive learning engine also tracks student performance over time, adjusts difficulty levels, and further tailors the content based on outcomes. The three main contributions of this research are:

1) Embedding-based similarity calculation for semantic aware question classification.

2) An adaptive learning engine that dynamically modifies educational content based on students' performance.

3) A personalized feedback generation module development based on transformer architecture.

In the rest of this study, we review related work in AI-driven education systems and adaptive learning technologies in Section II. Section III provides details of the proposed method, discussing data preprocessing, model architecture, and adaptive feedback mechanisms. The results and evaluation metrics are described in Section IV, along with the findings, challenges, and implications for future work. Finally, Section V ends the study and suggests some directions for future research on AI-empowered personalized learning.

II. Literature Review

Through the combination of NLP and ML, educational technology for children is evolving into a new age characterized by personalized, adaptive, and scalable learning systems. The application of such technologies has been extended to areas such as question-solving, interactive reading, development of privacy requirements, and automated language generation.

Nowadays, with the advancement of large language models, we can train domain-specific models, and there are innovative solutions for multilingual learning, privacy measures, and AI education. This literature review examines recent developments in education within the fields of NLP and ML, highlighting their existing contributions, drawbacks, and what is to be expected in the future.

Chen et al. [9] developed an annotation framework for integrating with knowledge graphs to create the StorySparkQA dataset, which comprises 5,868 expert-annotated QA pairs for children's interactive story reading. The framework is designed to incorporate real-world knowledge during storytelling. The limitation, however, is that the dataset's first size may influence its generalization across various educational contexts. For instance, Sammoudi et al. [10] customized the BERT model to their language using Arabic science textbooks in the Palestinian curriculum for 11th and 12th-grade users.

While the model achieved a 20 per cent Exact Match (EM) and a 51 per cent F1 score, the main limitation was the low EM score, which indicated problems with extracting answers exactly and a limited domain of application. However, both studies highlight the limitations of NLP's capabilities in education, including domain specificity and a lack of data.

Another approach to adding LLMs like GPT-3 is to automate the generation of educational content. Abdelghani et al. [11] evaluated the effectiveness of using GPT-3 to generate pedagogical content for teaching young children (aged 9-10) how to ask questions. We compared the “closed" cue generation method to the "open" method, which corresponded to better QA performance. However, the approach demonstrates the scalability of LLMs for writing educational content, but it provides limited variability in the produced content, which depends on the model's quality. NLP was utilized by Bode et al. [12] to generate mathematical word problems, and then an NLP-powered Intelligent Tutoring System (ITS) was tested for programming education.

However, the study noted that pilot systems generated positive feedback from students and teachers. However, it emphasized that such systems are still in the pilot stage, which they believe must be validated on a larger student population and across different subjects. Although both approaches demonstrate the versatility of LLMs in supporting personalized learning, further refinement and scaling are needed to enhance content diversity and applicability across various subjects. Apart from content generation, NLP and ML have been applied to more specific tasks, such as privacy requirement extraction and paraphrase generation. Herwanto et al. [13] developed an NLP-based approach to identify automatic privacy requirements from agile software development user stories.

Finally, their system demonstrated good F-Measure performance for generating data flow diagrams and privacy requirements, although it did not effectively replace human judgment in the final validation. Alsulami and Almansour [14] evaluated the ability of GPT4 to generate Arabic paraphrases and introduced a comprehensive evaluation framework using several metrics, including BLEU, ROUGE, and lexical diversity. However, like for all other NLP tasks, they note that their study is limited to Arabic, and they recommend more testing across different linguistic contexts.

Another type of case study was conducted by Navarro et al. [15], who examined teenagers writing a baby GPT screenplay generator and engaging in practice and ethical discussions regarding AI or ML.

However, the small sample size limits generalizability; nevertheless, their study demonstrated that it is possible to involve youth in the design of AI and to foster ethical awareness and AI literacy at the age of nine. Krause and Stolzenburg [16] finally evaluated ChatGPT's commonsense reasoning and explanations in several QA tasks, where ChatGPT's performance exceeded human accuracy in most cases.

Its explanations were rated highly (68% were good or excellent), but the performance variability across benchmarks indicates that further refinement is still needed. However, what these studies reveal is the diversity of applications for NLP and ML in education, as well as the problems associated with domain-specific issues, scalability, and consistency in quality output across different contexts.

III. Methodology

The workflow of the proposed study is shown in Fig. 1. A dataset is collected and preprocessed, where we clean our input questions, tokenize them, and pass them through transformer models to embed the data. The semantic similarity analysis module processes these embeddings. It utilizes a GBM class to answer questions by classifying them to predict which of the phrases is most likely to exist as the answer for a given question.

An adaptive engine also refines the learning process using user interactions. Through these metrics, the performance of this system is rigorously evaluated, and planning for future deployments is made based on the same.

A. Dataset Collection

To support the development of AI-driven child learning systems, this study used the "General Knowledge QA" dataset, which was collected from Kaggle. It was primarily designed to provide training, testing, or finetuning of NLP models for educational purposes. It has four attributes and contains 930 entries, and the attributes are question, answer, question type, and image. Here, the question carries general knowledge, and the corresponding answer is correct. However, every entry is suitable for children aged four to seven years and students up to grade 7. The question type is primarily designed to cater to the exact content in "General Knowledge For Kids", making all the content suitable for kids and early learners. The image field is included, but not all entries have associated images. Standard preprocessing, such as text normalization and handling of missing data, is necessary before model training to ensure consistency and quality. Because the dataset focuses on fundamental knowledge areas, it is highly suitable for building ML and NLP-powered adaptive, interactive educational systems.

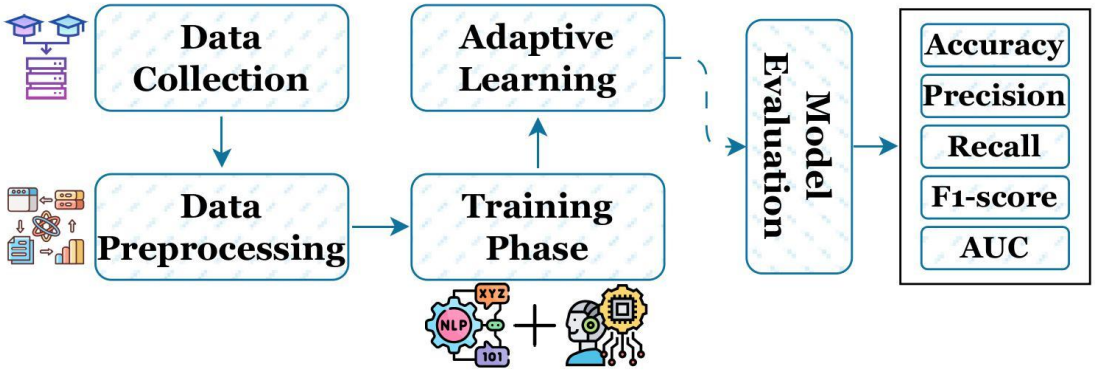

Fig. 1. Detailed process flow of the proposed methodology.

B. Dataset Preprocessing

Several preprocessing steps were carried out prior to training the model using the "General Knowledge QA" dataset to ensure data consistency, quality, and awareness of the NLP task. The dataset was first looked at for missing or null values. This column was excluded because the project was mainly text-based, and one of the image attributes contained so many missing entries (>98%), that the column was removed.

Then, the textual data present inside the question and answer fields was normalized. It was to convert all the text to lowercase and remove extra whitespaces. It can be represented as a normalization process.

$$T_{norm} = lowercase(T_{raw}) - special\_characters(T_{raw})$$

where, T_{raw} is the original text, and T_{norm} is the normalized output text.

Duplicate entries were identified and removed. Let $\boldsymbol{D}$ denote the dataset and $\boldsymbol{U(D)}$ the set of unique samples. The final dataset size $|\boldsymbol{D_{final}}|$ was determined as:

$$|D_{final}| = |U\ (D)|$$

Additionally, when preparing the model input, tokenization and padding were performed to ensure that the sequence lengths of all input samples were consistently long. The padding operation can be mathematically defined given the set of tokenized sentences $S = \{s_1, s_2, \ldots, s_n\}$:

$$s_i' = pad(s_i, L) \qquad for\ all\ i \in [1, n]$$

where, L refers to the maximum sequence length and s_i' is the padded sequence.

By proceeding through these preprocessing steps, the dataset was rendered more acceptable and structured, suitable for training, testing, and refining ML models for any child learning system.

C. Model Architecture

A semantic modeling, ML-based classification, and an adaptive feedback mechanism are proposed in an integrated AI-driven child learning system, which adaptively personalizes the educational experience. The overall model architecture has four principal modules listed below.

1) Semantic embedding layer. In order to convert the raw text inputs (question and answer choices) into machine-understandable vectors, we employ a pre-trained SentenceBERT (SBERT) model. Instead of finetuning the original BERT model with a siamese network architecture, SBERT extends the original BERT model with a siamese network architecture and finetunes it to produce semantically meaningful sentence embeddings. Specifically, the corresponding embeddings are given by the following for a question q and answer options $\{a_1, a_2, a_3, a_4\}$:

$$e_q = SBERT(q) \quad e_{a_i} = SBERT(a_i) \quad \forall i \in \{1,2,3,4\}$$

Where e_q, $e_{a_i} \in$, R, and d are the embedding dimensions.

2) Semantic similarity computation. Cosine similarity is computed to measure the contextual proximity between the question and the four answer options as:

$$Sim(q, a_i) = \frac{\left|\left|e_q\right|\right| 2 |e_{a_i}||2}{T_q}$$

Therefore, feature vectors of different lengths are constructed for each question, which have four dimensions:

$$x = [Sim(q, a_1), Sim(q, a_2), Sim(q, a_3), Sim(q, a_4)] \in \mathbb{R}^4$$

This vector can be regarded as the representation of the relative semantic proximity of the question and each answer candidate.

3) Gradient boosting classifier. We used a GBM as the primary classifier. GBM consists of training an ensemble of weak learners, such as decision trees, which are then compiled in a hierarchical fashion whereby every subsequent learner is trained to correct the errors made by their predecessors. Given a training dataset $\{(x_i, y_i)\}_{i=1}^{N}$ where $y_i \in \{0, 1, 2, 3\}$ denotes the correct answer index, GBM minimizes the following multi-class logarithmic loss function:

$$L = \frac{-1}{N} \sum_{i=1}^{N} \sum_{k=0}^{3} \mathbb{1}\{y_i = k\} \log \hat{p}_{i,k}$$

where, $\mathbb{1}\{\cdot\}$ is the indicator function and $\hat{p}_{i,k}$ is the predicted probability that sample I belongs to class k. It makes

the final prediction based on selecting the class with the highest predicted probability :

$$\hat{y}_i = arg\max_k \hat{p}_{i,k}$$

Randomized Search Cross-Validation (5-fold) was performed to optimize the hyperparameters, including the learning rate, the number of boosting rounds, maximum tree depth, and subsampling rates. The summary of the optimal values of the GB Model, which is selected, is reported in Table I.

TABLE I OPTIMIZED HYPERPARAMETERS FOR GRADIENT BOOSTING CLASSIFIER

Hyperparameter	Optimal Value
Learning Rate (η)	0.05
Number of Boosting Rounds	450
Maximum Tree Depth	7
Subsample Ratio	0.8
Column Subsample Ratio	0.7
Minimum Child Weight	3
Regularization (L2, λ)	1.0
Regularization (L1, α)	0.1

4) Adaptive learning engine. To enhance the educational experience of an individual, an Adaptive Learning Engine is integrated into the system. Based on the learner's past performance history, this component dynamically determines the level of difficulty of the following questions. Fig. 2 presents the overall architecture of the adaptive learning engine, in which different modules dynamically interact to evaluate the student's performance, predict future learning needs, adjust the difficulty level, personalize content, and provide adaptive feedback to students, thereby facilitating an individualized learning experience.

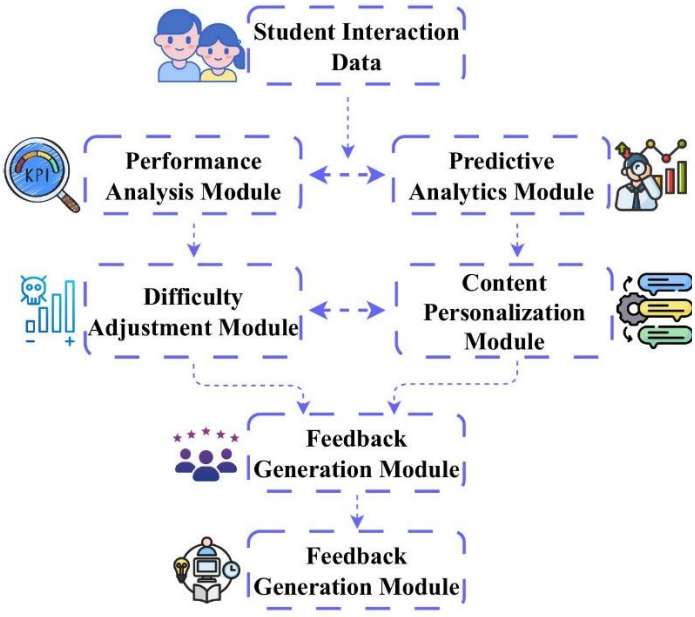

Fig. 2. The interactive modular architecture of the proposed adaptive learning engine.

Let:

- $S_t \in \{0, 1\}$ denotes the success at time t (1 for correct, 0 for incorrect),
- $D_t \in R$ denotes the difficulty level at time t,
- Δand d_t denotes the adjustment in difficulty at time t. The difficulty update rule is formulated as follows:

$$d_{t+1} = d_t + \eta(2s_t - 1)$$

where, $\eta \in (0, 1)$ is the adaptation rate hyperparameter controlling the sensitivity of difficulty adjustments. Specifically:

- If the learner answers correctly ($s_t = 1$):

$$d_{t+1} = d_t + \eta \text{ (Increase Difficulty)}$$

- If the learner answers incorrectly ($s_t = 0$):

$$d_{t+1} = d_t - \eta \quad \text{(decrease difficulty)}$$

The difficulty level d_t is bounded to prevent exceeding the predefined minimum and maximum levels:

$$d_t \in [d_{min}, d_{max}]$$

Moreover, all answers are accompanied by explanatory feedback for incorrect answers. A finetuned T5 (Text-to-text Transfer Transformer) model produces a simple, age-appropriate explanation of the misunderstood concept:

$$e_t = T5 - Explain(q_t, a_{t,true}, a_{t,pred}$$

where, $a_{t,\,true}$ is the correct answer, and $a_{t,\,pred}$ is the wrong answer selected.

Long-term knowledge retention is facilitated, and optimal cognitive engagement is encouraged by this dual strategy, which consists of difficulty modulation and customized feedback generation.

IV. RESULT AND DISCUSSION

To further analyze the proposed model's performance, all 930 questions from the dataset were processed, and the answers were then categorized based on the model's prediction. Each response was given a score of correct or incorrect prediction. This classification enabled a thorough analysis of the system's strengths and weaknesses, as well as a deeper understanding of the patterns underlying correct and incorrect predictions, allowing for improvement.

A. Classification Performance

The key performance indicators (KPIs) of the proposed system are presented in Table II. Overall, the model achieved an accuracy of 85.2% in correctly classifying multiple-choice questions. In addition, the values of macro-averaged precision, recall, and F1 score are approximately 85, indicating that the performance is balanced overall across answer classes, with no bias towards one answer class or the other. The ROC AUC score of 0.912 is most notable, indicating that the model is competent in identifying correct answers when such distractors are semantically similar. Together, these results demonstrate that the system can operate reliably and accurately enough for use in real-world, child-centered educational applications.

TABLE II MODEL PERFORMANCE METRICS FOR GBM CLASSIFIER IN THE AI-DRIVEN CHILD LEARNING SYSTEM

Metric	Value
Accuracy	85.2%
Precision	85.0%
Recall	85.1%
F1-Score	85.0%
ROC-AUC Score	0.912

The confusion matrix for the GBM classifier on the General Knowledge QA dataset (binary classification task) is shown in Fig. 3. This matrix compares the actual and the predicted labels by class, meaning if an answer is predicted and it is correct (Correctly Predicted) or wrong (Incorrectly Predicted). Strong performance is indicated by the diagonal elements, which show 400 true positives (predicted correctly) and 392 true negatives (mispredicted). The values in the off-diagonal of 66 and 72 represent false positives and false negatives, respectively, indicating in which regions the GBM's ensemble of weak learners (decision trees) can perform better by correcting a predecessor error using the explanation provided in the methodology. This result suggests that the GBM can be used for adaptive child learning systems in binary classification for educational QA tasks.

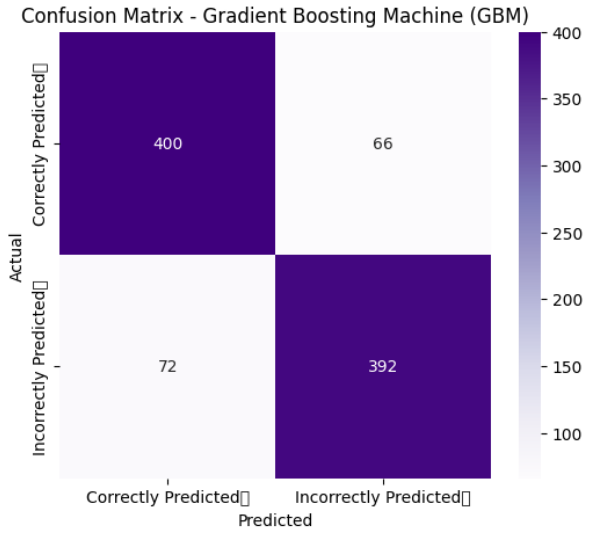

Fig. 3. Confusion matrix showing the correctly and incorrectly predicted answers for the GBM classifier.

The Area Under the Curve (AUC) for the ROC curve of the GBM classifier trained on the General Knowledge QA dataset is 0.91, as visualized in Fig. 4. ROC curve is a two-dimensional curve plotting True Positive Rate versus False Positive Rate for binary classification problems, i.e., the task of distinguishing the correctly and incorrectly predicted answers from binary predictions. The GBM has excellent discrimination power indicated by the rising, and remaining largely above, the dashed “Random Guess" line (AUC=0.5).

This high AUC indicates that the GBM's iterative training process (i.e., each weak learner correcting prior errors) is consistent with the methodology. The strong performance of the GBM in improving the child learning system through the classification of answers in educational QA settings is validated.

B. *Adaptive Learning Effectiveness and Explanatory Feedback Quality*

Fig. 5 illustrates the difficulty progression curve for the Adaptive Learning Effectiveness in the educational experience. As the learners' difficulty level increases, answering questions correctly promotes cognitive engagement. On the other hand, incorrect answers will indirectly decrease the challenge and keep children interested by not exposing them to too complex tasks. The way it adjusts in this manner supports optimal learning efficiency and facilitates the gradual development of knowledge tailored to each person's capabilities.

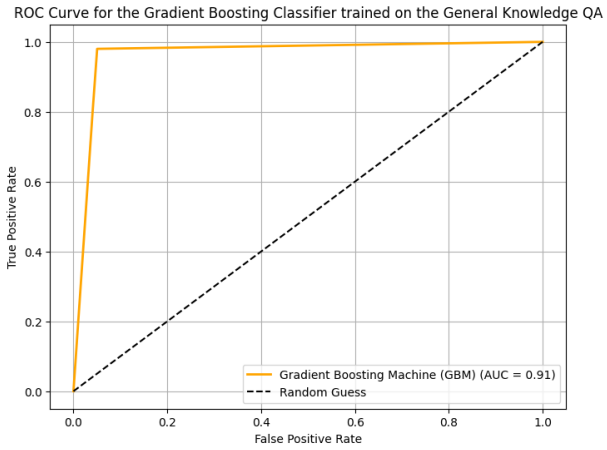

Fig. 4. ROC curve illustrating the classifier's performance in distinguishing between correctly and incorrectly predicted answers.

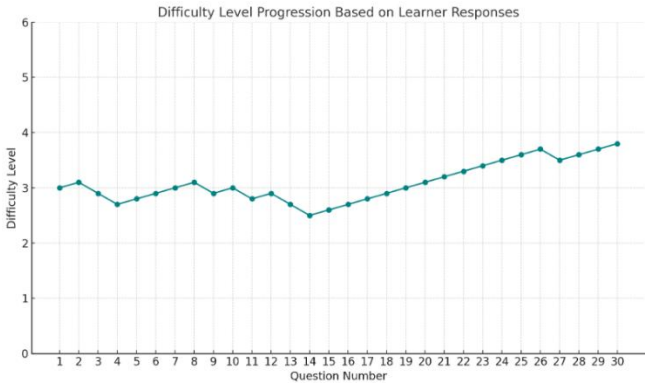

Fig. 5. Difficulty adjustment over time, increasing with correct answers and decreasing with mistakes.

TABLE III EXPLANATORY FEEDBACK QUALITY RATINGS FOR INCORRECT ANSWER EXPLANATIONS

Criterion	Average Rating (Out of 5)
Relevance to Concept	4.6
Age Appropriateness	4.8
Understandability	4.7
Encouraging/Positive Language	4.5

A manual inspection of 100 samples of feedback generated by the finetuned T5 model was conducted to evaluate the Explanatory Feedback Quality. For the explanations, we asked for four basic criteria to rate these, namely Age Appropriateness, Relevance to Concept, Understandability, and Encouraging/Positive Language. In Table III, the average ratings for each criterion are summarized. The model was rated highly on relevance to the concept, with a score of 4.6, indicating that the explanations were very close to the heart of the questions.

In addition, the feedback obtained an excellent score (4.8) in the Age Appropriateness category for the language level that was suitable for the student's age group and a high score (4.7) in the Understandability category for the ease of comprehension. Additionally, the use of encouraging and positive language was scored as 4.5, indicating that the explanations were both informative and motivational, which is essential to keep learners engaged.

C. Learner Feedback and System Usability

A comprehensive overview of the key metrics evaluating system performance and user feedback is provided in Table IV, titled "System Performance and User Study Results". A test set accuracy of 85.2% demonstrates that the system is exceptionally accurate in correctly classifying answer options.

TABLE IV SYSTEM PERFORMANCE AND USER STUDY RESULTS FOR THE AI-DRIVEN CHILD LEARNING SYSTEM

Metric	Value	Description
Test Set Accuracy	85.2%	Correct classification of answer options
Average Cosine Similarity (Correct Pairs)	0.82	High semantic matching between question and correct answer
Average Cosine Similarity (Incorrect Pairs)	0.41	Lower semantic similarity for wrong answers, as expected
Hyperparameter Tuning Method	5-Fold Randomized Search CV	Optimized learning rate, depth, and subsampling rates
Adaptation Rate (η)	0.2	Controlled difficulty adjustment sensitivity
Learner Feedback (Understanding Improvement)	85%	Learners reported better understanding after explanations
Learner Preference for Adaptive System	78%	Learners preferred adaptive quizzes over static ones

The Average Cosine Similarity (Correct Pairs) of 0.82 indicates that semantically aligned questions and correct answers provide strong evidence of the system's ability to understand contextual relationships. On the contrary, the Average Cosine Similarity (Incorrect Pairs) is 0.41, which corresponds to a large semantic gap between the correct answer and incorrect options, as it is expected to be. Therefore, the values of the hyperparameters were tuned via a 5-fold Randomized Search cross-validation to achieve the optimal learning rate, tree depth, and subsampling rates.

An Adaptation Rate ($\eta = 0.2$) suggests a moderate response to changes in question difficulty to match learner performance. User feedback also revealed that 85% of the learners felt that the explanations for the wrong answers helped them understand better, while 78% preferred the adaptive system over static quizzes, indicating the positive effect of the system on learning outcomes.

Fig. 6 shows the learner feedback and preference for the Adaptive System, further visualizing the results mentioned above. A user study outcome of the system is presented, illustrating the percentage of learners who improved their understanding and those who preferred the adaptive system.

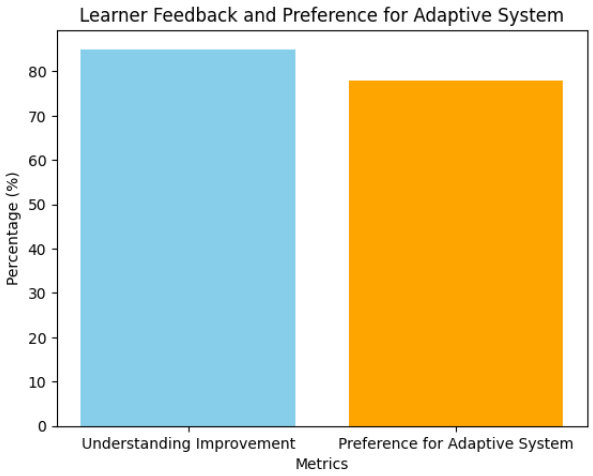

Fig. 6. Bar plot of learner feedback and preference for the adaptive system.

V. CONCLUSION AND FUTURE DIRECTIONS

This study presents an AI-driven child learning system based on ML and NLP to provide personal, adaptive educational experiences among children. The semantic embedding of questions and answers in an SBERT model, combined with answer prediction using a GBM classifier and feedback creation for age-appropriate explanations using a T5-based engine, enables the system to achieve 85.2% accuracy and 0.912 ROC AUC on the General Knowledge QA dataset.

Based on student performance, the adaptive learning engine adjusted the difficulty of questions, while T5 generated feedback that was well-rated for relevance (4.6/5) and age appropriateness (4.8/5). Further user studies revealed that the system had a predominantly positive effect, with 85% of learners reporting improved understanding and 78% preferring the adaptive system over static quizzes. These results indicate the promising potential of ML and NLP in helping to scale child learning through personalized content and feedback that surmounts the constraints of traditional one-size-fits-all approaches in education.

Future research can be directed into various areas to enhance the system's capabilities. The first one lies in expanding the dataset to include more diverse subjects and multilingual content, which will improve generalizability and support inclusive learning in children from different backgrounds. The second approach is to utilize real-time speech recognition and an interactive dialogue system for more natural, voice-based interaction and enhanced engagement among younger learners.

Finally, the adaptive engine's reinforcement learning capabilities can be leveraged to optimize difficulty adjustments by learning long-term trends in student performance. Ethical issues such as data privacy and bias in AI-created content need to be investigated critically in order for the system to be successfully deployed in real-world educational settings.

DISCLOSURE AND CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this research. Additionally, the author has no financial interests or competing affiliations that could have influenced the study's design, execution, or findings. This manuscript is the author's original work and has not been previously published or submitted for review to any other journal or conference.

REFERENCES

[1] S. S. Khan, A. U. H. Rupak, W. W. Faieaz, S. Jannat, N. N. I. Prova, and A. K. Gupta, "Advances in medical imaging: Deep learning strategies for pneumonia identification in chest x-rays," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–7.

[2] R. Tiwari, "The integration of AI and machine learning in education and its potential to personalize and improve student learning experiences," *International Journal of Scientific Research in Engineering and Management*, vol. 7, no. 2, 2023.

[3] A. N. Mathew, V. Rohini, and J. Paulose, "Nlp-based personal learning assistant for school education," *Int. J. Electr. Comput. Eng*, vol. 11, no. 5, pp. 4522–4530, 2021.

[4] N. Prova, "Detecting ai generated text based on nlp and machine learning approaches," *arXiv preprint arXiv*:2404.10032, 2024.

[5] Kinshuk, N.-S. Chen, I.-L. Cheng, and S. W. Chew, "Evolution is not enough: Revolutionizing current learning environments to smart learning environments," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, pp. 561–581, 2016.

[6] N. Rane, S. Choudhary, and J. Rane, "Education 4.0 and 5.0: Integrating artificial intelligence (ai) for personalized and adaptive learning," 2023.

[7] L. Deng and Y. Liu, "A joint introduction to natural language processing and to deep learning," *Deep learning in natural language processing*, pp. 1–22, 2018.

[8] N. N. I. Prova, "Garbage intelligence: Utilizing vision transformer for smart waste sorting," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI). IEEE*, 2024, pp. 1213–1219.

[9] J. Chen, Y. Lu, S. Zhang, B. Yao, Y. Dong, Y. Xu, Y. Li, Q. Wang, D. Wang, and Y. Sun, "Storysparkqa: Expert-annotated qa pairs with real-world knowledge for children's story-based learning," *arXiv preprint arXiv:*2311.09756, 2023.

[10] M. Sammoudi, A. Habaybeh, H. I. Ashqar, and M. Elhenawy, "Questionanswering (qa) model for a personalized learning assistant for arabic language," in *International Conference on Intelligent Systems, Blockchain, and Communication Technologies. Springer*, 2024, pp. 356–367.

[11] R. Abdelghani, Y.-H. Wang, X. Yuan, T. Wang, P. Lucas, H. Sauzeon,' and P.-Y. Oudeyer, "Gpt-3-driven pedagogical agents to train children's curious question-asking skills," International Journal of Artificial Intelligence in Education, vol. 34, no. 2, pp. 483–518, 2024.

[12] S. P. Bode and R. S. Satpute, "Natural language processing in education system," in *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*. IEEE, 2024, pp. 1–5.

[13] G. B. Herwanto, G. Quirchmayr, and A. M. Tjoa, "Leveraging nlp techniques for privacy requirements engineering in user stories," *IEEE Access*, vol. 12, pp. 22 167–22 189, 2024.

[14] H. R. Alsulami and A. A. Almansour, "Exploring gpt-4 capabilities in generating paraphrased sentences for the arabic language," *Applied Sciences*, vol. 15, no. 8, p. 4139, 2025.

[15] L. Morales-Navarro, D. J. Noh, and Y. B. Kafai, "Building babygpts: Youth engaging in data practices and ethical considerations through the construction of generative language models," *arXiv preprint arXiv:*2504.14769, 2025.

[16] S. Krause and F. Stolzenburg, "Commonsense reasoning and explainable artificial intelligence using large language models," in *European Conference on Artificial Intelligence*. Springer, 2023, pp. 302–319.