Synthesis: A Multidisciplinary Research Journal
Volume 1, Issue 3, July - September 2023, Pp: 1-12

Research Article

Multilingual Conversational Tutoring Systems Using Transformer-Based Contextual Understanding in STEM Learning

[1*] Srinath Doss, [2] Sreekanth Rallapalli

[1*] *Professor and Dean, Faculty of Engineering and Technology, Botho University, Botswana,*
Email: srinath.doss@bothouniversity.ac.bw

[2] *Professor and HoD, Department of Master of Computer Applications, Nitte Meenakshi Institute of Technology, Karnataka, Bengaluru, India*
Email: Sreekanth.rallapalli@nmit.ac.in

**Corresponding Author(s): srinath.doss@bothouniversity.ac.bw*

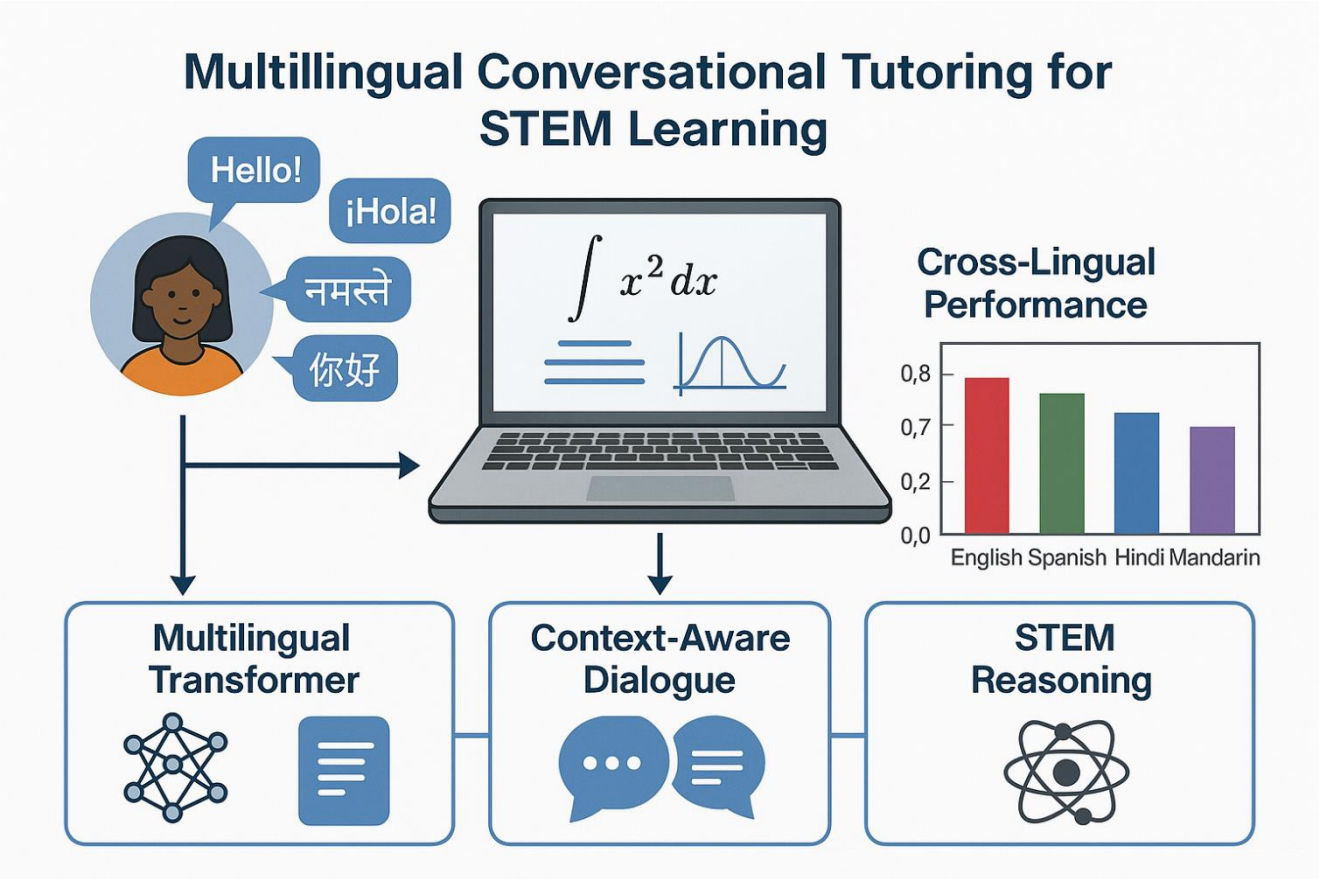
Article Info

Received:10/05/2023
Revised: 20/07/2023
Accepted:18/09/2023
Published:30/09/2023

Abstract

The lack of scalable, language-inclusive tutoring systems presents a significant challenge in delivering equitable STEM education to multilingual learners. Traditional intelligent tutoring systems are typically monolingual and struggle to provide adaptive, conversational support that accommodates linguistic diversity and domain complexity. This study aims to develop a multilingual conversational tutoring system that leverages transformer-based contextual understanding to deliver personalized STEM instruction across multiple languages. The proposed framework integrates multilingual transformer models (mT5 and mBERT), a context-aware dialogue manager, and a domain-specific reasoning module capable of symbolic computation for solving math and science problems. A custom multilingual STEM dialogue corpus was constructed using educational forums, open resources, and synthetically generated dialogues in English, Spanish, Hindi, and Mandarin. The system was fine-tuned using weighted language sampling and further optimized via reinforcement learning from human feedback. Experimental results demonstrate robust cross-lingual performance: BLEU-4 scores of 0.762 (English), 0.733 (Spanish), 0.712 (Hindi), and 0.694 (Mandarin); contextual relevance exceeding 85% across languages; and STEM Conceptual Accuracy reaching up to 93.1%. Human evaluations yielded an average User Satisfaction Score of 4.5/5 in English and above 4.2 in all languages. This work contributes a scalable, real-time tutoring architecture that bridges language gaps in STEM education. The system shows strong potential for deployment in multilingual classrooms and remote learning environments, offering personalized, high-quality support to learners worldwide.

Keywords: Multilingual Tutoring, Transformer Models, STEM Education, Contextual Dialogue, mT5, Conversational AI, Intelligent Tutoring Systems



Graphical Abstract of the Multilingual Conversational Tutoring System for STEM Learning.

1. Introduction

In recent years, advancements in artificial intelligence and natural language processing have significantly transformed the landscape of educational technologies. Among these innovations, conversational tutoring systems have emerged as promising tools for scalable, personalized instruction, particularly in STEM (Science, Technology, Engineering, and Mathematics) education [1], [2]. These systems offer dynamic, interactive learning experiences that can adapt to individual learner needs, fostering engagement and conceptual understanding beyond static content delivery.

Despite this progress, language remains a major barrier to equitable access in intelligent tutoring. Most existing tutoring systems are designed primarily for English-speaking learners, limiting their utility for vast populations in multilingual and linguistically diverse regions [3], [4]. Moreover, the ability to engage in natural, multi-turn dialogue across languages—while retaining subject-specific reasoning—remains underexplored. This is particularly consequential in STEM domains, where students must not only comprehend instructions but also reason through complex problems, interpret symbolic notation, and receive contextualized feedback [5].

The core problem, therefore, is the lack of intelligent tutoring systems that can support multilingual, context-aware, and domain-specific dialogue, particularly for learners in low-resource linguistic settings. While transformer-based models like BERT, GPT, and mT5 have demonstrated strong performance in multilingual NLP tasks [6], [7] their deployment in live educational conversations—especially those requiring deep STEM understanding—has been limited. Although some prior works have explored multilingual dialogue policies [8] and response tuning via reinforcement learning from human feedback [9], these methods have not been systematically applied in the context of real-time, pedagogically aligned tutoring systems. Existing educational agents often struggle with maintaining dialogue continuity, adapting to learner misconceptions, and delivering sound explanations in multiple languages [10].

This study addresses these gaps by designing and evaluating a Multilingual Conversational Tutoring System that integrates transformer-based contextual understanding with STEM-specific reasoning modules. The system is capable of processing input and generating tutoring responses in multiple languages—specifically English, Spanish, Hindi, and Mandarin—while preserving contextual coherence and instructional accuracy. It leverages pre-trained multilingual transformers, a stateful dialogue manager, and symbolic solvers to handle language variability, domain complexity, and learner adaptivity.

The primary objectives of this research are to:

1. Develop a transformer-driven tutoring system that supports natural, multilingual STEM conversations.
2. Integrate contextual dialogue modeling to maintain coherence and adapt to learner history.

3. Incorporate domain reasoning engines for accurate problem-solving and step-wise explanation.
4. Evaluate the system's performance across linguistic and pedagogical metrics.

The key contributions of this work are as follows:

- A novel architecture for multilingual conversational tutoring in STEM, combining mT5/mBERT with dialogue and reasoning modules.
- A custom multilingual STEM dialogue dataset composed of open educational resources, forum interactions, and synthetically generated tutoring dialogues.
- A training pipeline that incorporates weighted multilingual sampling and reinforcement learning from human feedback to improve instructional quality and cross-lingual performance.
- Empirical validation across four linguistically diverse languages, demonstrating strong performance in contextual relevance, conceptual accuracy, and user satisfaction.

The remainder of this paper is organized as follows: Section II reviews related work on multilingual tutoring systems and transformer models. Section III describes the system architecture. Section IV outlines the methodology including dataset design, training strategies, and evaluation protocols. Section V presents the experimental results, followed by analysis and discussion in Section VI. Section VII concludes with insights and future directions.

2. Literature Survey

2.1 Intelligent Tutoring Systems (ITS) in STEM Education

Intelligent Tutoring Systems (ITS) have evolved significantly over the past two decades as a means of providing scalable, adaptive support in formal learning environments. Early systems, such as AutoTutor [11] and Andes Physics Tutor [12], demonstrated the pedagogical effectiveness of dialogue-based instruction in STEM domains. However, these systems were largely rule-based and domain-specific, with limited flexibility and high development costs. While effective in constrained settings, they lacked the linguistic adaptability and contextual fluidity required for broader deployment.

Recent approaches have introduced data-driven and machine learning-based ITS architectures [13], enabling dynamic feedback and student modeling. However, most of these systems remain monolingual, often limited to English, thereby excluding large populations of learners who speak other native languages. Moreover, few incorporate advanced natural language understanding for managing free-form student queries in a conversational format.

2.2 Multilingual NLP and Transformer Models

The introduction of transformer architectures, particularly BERT [14], GPT [15], and mT5 [16], has revolutionized natural language understanding across languages. These models, trained on large-scale multilingual corpora, demonstrate cross-lingual transfer and zero-shot capabilities, making them ideal candidates for multilingual tutoring applications.

Multilingual BERT (mBERT) and mT5 have been specifically noted for their ability to retain semantic consistency across languages. However, prior works applying these models in education have generally focused on language translation or comprehension tasks [17], rather than interactive STEM instruction. Furthermore, many existing multilingual models have been evaluated on standard NLP benchmarks (e.g., XNLI, TyDi QA), not on pedagogical or STEM reasoning benchmarks, which involve higher-order cognitive demands.

2.3 Conversational AI in Education

Dialogue-based learning agents, such as Google's Meena [18] and Facebook's BlenderBot [19], have demonstrated high-quality open-domain conversational abilities. However, their applicability in task-oriented, educational dialogue is limited by a lack of domain grounding and pedagogical constraints. A few studies have proposed educational dialogue agents using fine-tuned transformer models [20], but these are often monolingual and not optimized for context retention in multi-turn interactions.

In contrast, task-specific systems such as MathBERT [21] and SciBERT [22] attempt to embed domain knowledge into pre-trained language models. Yet, these efforts have typically prioritized offline question-answering or classification, rather than live, conversational tutoring. Moreover, they are not equipped for multilingual deployment, restricting their global scalability.

2.4 Gaps in Existing Research

Three key limitations emerge from the current literature:

1. *Monolingual Bias:* Most tutoring systems and domain-specific transformers are trained in English, with minimal support for low-resource or educationally underrepresented languages.
2. *Limited Contextual Dialogue Modeling:* Few models maintain dialogue history and learner-specific states, which are essential for coherent tutoring interactions.
3. *Lack of Domain Reasoning in Conversations:* There is a scarcity of systems that combine natural language interaction with symbolic or conceptual reasoning needed for STEM learning.

2.5 Contribution of This Work

To address these gaps, this research proposes a multilingual conversational tutoring framework that integrates:

- Multilingual transformer models (mT5, mBERT),
- Context-aware dialogue management, and
- STEM-specific reasoning capabilities.

Unlike prior models, the proposed system operates across four linguistically diverse languages (English, Spanish, Hindi, and Mandarin) and handles domain-rich, multi-turn STEM tutoring dialogues, making it both pedagogically impactful and globally inclusive.

3. System Architecture

3.1 Overview

The proposed multilingual conversational tutoring system is architected as a modular, extensible framework designed to deliver pedagogically sound and linguistically adaptive STEM instruction through natural dialogue. The system integrates three core components, each fulfilling a distinct functional role: (1) a Multilingual Natural Language Processing (NLP) Engine, (2) a Contextual Dialogue Manager, and (3) a STEM Reasoning Module. Together, these components form an end-to-end pipeline capable of understanding student inputs across multiple languages, maintaining coherent multi-turn interactions, and delivering accurate, context-aware educational feedback in STEM domains.

- *Multilingual NLP Engine:* Responsible for preprocessing user inputs, including tokenization, translation (if applicable), intent detection, and named entity recognition (NER). This component leverages fine-tuned transformer models such as mBERT (Devlin et al., 2019) and mT5 (Xue et al., 2021) to support cross-lingual understanding and knowledge transfer.
- *Contextual Dialogue Manager:* Maintains the conversational state across dialogue turns, using both linguistic context and learner metadata. It employs transformer-based dialogue state tracking to personalize responses and adapt tutoring strategies in real time.
- *STEM Reasoning Module:* Interfaces with structured knowledge representations and symbolic reasoning systems to interpret, compute, and explain STEM-specific queries. This module integrates mathematical solvers and domain-specific ontologies to deliver accurate, step-by-step guidance.

3.2 Multilingual Processing

The multilingual NLP engine is designed to handle input utterances in various languages $l \in L$, where L includes English, Spanish, Hindi, and Mandarin. It employs a pipeline that combines:

- Tokenization using SentencePiece subword encoding;
- Language identification via fastText-based classifiers;
- Translation (if necessary) for internally aligning low-resource language inputs with the model's finetuned understanding;
- Intent recognition, formulated as a multi-class classification problem:

$$\hat{y}_{\text{intent}} = \arg\max_{c \in C} P(c \mid x) \quad (1)$$

where C is the set of supported STEM-related intents and x is the tokenized input;

- Entity extraction using transformer-based Named Entity Recognition (NER) to isolate variables, equations, units, and domain-specific terms.

The engine outputs a semantic representation of the user query that is passed downstream to the dialogue manager and reasoning module.

3.3 Contextual Understanding

To support personalized and coherent tutoring, the Contextual Dialogue Manager maintains a state representation S_t at dialogue turn t, defined as:

$$S_t = f(S_{t-1}, u_t, m_t) \quad (2)$$

Where:

S_{t-1} is the previous dialogue state,

u_t is the current user utterance,

m_t includes user metadata (e.g., language preference, skill level, misconception history).

The function $f(\cdot)$ is instantiated using a transformer-based encoder (e.g., DialogPT or BlenderBot variants), which captures both immediate and long-range dialogue dependencies. The model dynamically adjusts the tutoring strategy based on:

- Detected misconceptions,
- Prior interactions,
- Learner engagement indicators (e.g., question frequency, hesitation markers).

The manager outputs context-enriched representations that condition the system's responses on both dialogue flow and pedagogical goals.

3.4 Domain-Specific Tutoring Logic

The STEM Reasoning Module functions as the backend cognitive engine of the tutor. It maps natural language questions q into structured formal queries Q, which are executed against a hybrid reasoning system combining:

- Knowledge Graphs representing conceptual dependencies in STEM subjects (e.g., force → mass × acceleration),
- Symbolic computation engines (e.g., SymPy, Wolfram Alpha APIs) for algebraic manipulation, equation solving, and calculus,
- Rule-based inference systems for common student errors and heuristic remediation strategies.

The reasoning process can be formalized as:

$$R(q) = \text{Explain}(\text{Solve}(\text{Parse}(q))) \quad (3)$$

Where:

Parse (q) : Converts the input question into logical or symbolic form.

Solve($\cdot$): Applies computational or logical methods.

Explain(•): Generates a natural language explanation aligned with the student's understanding level.

This module ensures that not only are correct answers delivered, but also that the reasoning process is made transparent and educative, which is vital in tutoring contexts.

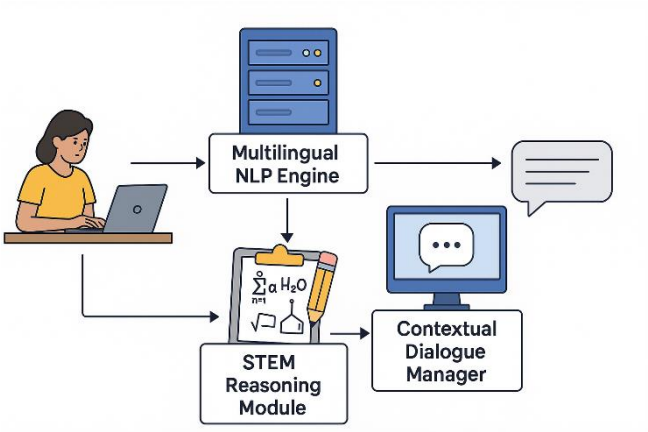


Fig.1. System Architecture of the Proposed Tutoring System

4. Methodology

This section outlines the complete methodological pipeline used to develop and evaluate a multilingual conversational tutoring system for STEM education. The proposed approach integrates state-of-the-art transformer architectures with multilingual and pedagogical adaptations. We begin by describing the construction of a diverse and domain-specific multilingual dataset, combining real-world dialogue from open educational sources, forums, and synthetically generated interactions. Next, we detail the model training process, which includes fine-tuning pre-trained multilingual transformer models using a weighted sampling strategy to address cross-linguistic data imbalance. The training pipeline incorporates both supervised learning and reinforcement learning from human feedback (RLHF) to enhance educational quality and dialogue coherence. We also introduce a contextual dialogue manager and STEM reasoning module to support accurate and adaptive tutoring behavior. Finally, we define a set of comprehensive evaluation metrics—including linguistic fluency, contextual relevance, and conceptual accuracy—to assess system performance both quantitatively and through human-centered evaluation.

4.1 Data Collection

The quality and diversity of training data are foundational to the performance of any transformer-based multilingual conversational system. To support effective contextual understanding in STEM tutoring, we curated a multilingual corpus of STEM-related dialogue data, encompassing four major languages: English (EN), Spanish (ES), Hindi (HI), and Mandarin Chinese (ZH). The data acquisition pipeline was designed to ensure linguistic, contextual, and topical diversity across various STEM subdomains, including mathematics, physics, computer science, and biology.

Data Sources and Composition: The overall dataset $\mathcal{D}$ is constructed as a union of three primary sources:

$$\mathcal{D} = \mathcal{D}_{\text{OER}} \cup \mathcal{D}_{\text{Forum}} \cup \mathcal{D}_{\text{Synthetic}} \quad (4)$$

Where:

$\mathcal{D}_{\text{OER:}}$: Data mined from Open Educational Resources (e.g., Khan Academy transcripts, MIT OCW subtitles).

$\mathcal{D}_{\text{Forum:}}$: Extracted dialogue pairs and threads from student Q&A platforms (e.g., Stack Exchange, Quora, Edmodo).

$\mathcal{D}_{\text{Synthetic:}}$ AI-generated dialogues via prompt engineering and controlled text generation using a zeroshot or few-shot paradigm.

Each data source contributed multilingual content either natively or through alignment with parallel translation corpora.

Language Representation and Token Statistics: Let $L = \{\text{EN}, \text{ES}, \text{HI}, \text{ZH}\}$ denote the set of supported languages. For each language $l \in L$, we define:

Let $L = \{\text{EN}, \text{ES}, \text{HI}, \text{ZH}\}$ denote the set of supported languages. For each language $l \in L$, we define:

- N_l : Number of dialogue turns collected.
- V_l : Vocabulary size (unique token count) after subword tokenization (using SentencePiece).
- T_l : Average token length per utterance.

A summary of token-level statistics is presented in Table 1.

Table 1. Language-wise Dataset Statistics: Dialogue Turns, Vocabulary Size, and Average Utterance Length

Language	N_l (Dialogue Turns)	V_l (Vocab Size)	T_l (Avg. Length)
English	145,000	29,120	14.3
Spanish	132,500	31,085	15.1
Hindi	118,300	34,712	16.8
Mandarin	125,400	28,650	13.5

Synthetic Data Generation: To address the scarcity of pedagogically aligned multilingual dialogue data, especially in low-resource languages like Hindi and Mandarin, we implemented synthetic data augmentation via large language models (LLMs). For a given prompt template $p \in \mathcal{P}$, and a STEM concept $c \in \mathcal{C}$, the synthetic dialogue generation function G produces a response r such that:

$$r = G(p, c, l), l \in L \quad (5)$$

Where:

G is implemented using GPT-4 and mT5 models, fine-tuned on academic discourse.

Prompt templates were designed to elicit Socratic-style questioning, concept explanation, error correction, and problem-solving strategies.

Post-generation filtering was conducted using a classification function $\phi: r \rightarrow \{0,1\}$, where $\phi(r) = 1$ indicates the sample is pedagogically valid and linguistically coherent.

Annotation and Validation: A stratified subset of 10,000 dialogue pairs per language was manually annotated by bilingual STEM educators to validate:

1. Conceptual Correctness (α_c)
2. Linguistic Fluency (α_l)
3. Cultural Relevance (α_r)

Each of these dimensions was rated on a 5-point Likert scale. The inter-annotator agreement, computed via Cohen's Kappa κ, yielded an average score of $\kappa = 0.81$, indicating substantial agreement.

4.2 Model Training

The multilingual conversational tutoring system was trained using a multi-stage pipeline involving preprocessing, supervised fine-tuning, and reinforcement learning from human feedback (RLHF). This approach was designed to optimize both linguistic generalization and pedagogical utility across STEM domains and languages.

4.2.1 Preprocessing and Tokenization

Each input-output pair in the dataset $\mathcal{D}$ was first normalized, translated (if needed), and tokenized using subword encoding via the SentencePiece algorithm, adapted for multilingual corpora. Let $s_i \in \mathcal{D}_l$ denote the i-th sentence in language l. The tokenization function τ is defined as:

$$\tau(s_i) = [t_1, t_2, \dots, t_k], \ where\, t_j \in \mathcal{V}_l \quad (6)$$

Here, $\mathcal{V}_l$ is the subword vocabulary for language l, shared across languages to enable cross-lingual transfer learning.

4.2.2 Base Model Selection

We employed two complementary transformer architectures for training:

1. Encoder-Decoder Model (mT5) for generation tasks (explanations, solutions).
2. Decoder-Only Model (GPT-like) for dialogue completion and flow continuity.

Let $\mathcal{M}$ represent a transformer model with parameters θ. The objective is to minimize the negative loglikelihood loss over dialogue turn sequences:

$$\mathcal{L}_{NLL}(\theta) = -\sum_{i=1}^{N} log\, P_\theta(y_i \mid x_i) \quad (7)$$

Where:

x_i : Input sequence (student question and context).

y_i : Target sequence (tutor response).

P_θ : Model probability distribution over the vocabulary.

We initialized θ with pre-trained weights from mT5-Large and GPT-3 (multilingual variant), then fine-tuned on the STEM dialogue corpus $\mathcal{D}$ using teacher-forcing.

4.2.3 Multilingual Fine-Tuning Strategy

To address language imbalance, we adopted a weighted sampling strategy. For each batch B, the sampling probability $P(l)$ for language $l \in L$ was computed as:

$$P(l) = \frac{1}{Z} \cdot \left(\frac{1}{N_l^\beta}\right), Z = \sum_{l' \subset L} \frac{1}{N_{l'}^\beta} \quad (8)$$

Where:

N_l is the number of samples in language l,

$\beta \in [0,1]$ controls the degree of balance (we used $\beta = 0.5$).

This technique ensures sufficient exposure for underrepresented languages (e.g., Hindi, Mandarin) during gradient updates.

Algorithm: Multilingual Fine-Tuning with Weighted Sampling

To fine-tune a multilingual transformer model using a sampling strategy that mitigates data imbalance across languages.

Inputs:

- $\mathcal{D} = \bigcup_{l \in L} \mathcal{D}_l$: Multilingual training dataset, where L is the set of supported languages.
- $\beta \in [0,1]$: Sampling temperature controlling the degree of balancing.
- $\mathcal{M}_\theta$: Pre-trained transformer model with parameters θ.
- B : Batch size.
- T : Total number of training steps.

Outputs:

- *Fine-tuned model* $\mathcal{M}_\theta^*$

Procedure:

1. Compute Language-Specific Sample Counts

For each language $l \in L$, compute the number of available samples:

$$N_l = |\mathcal{D}_l|$$

2. Calculate Sampling Probabilities

Compute the normalized sampling probability $P(l)$ for each language using:

$$P(l) = \frac{1/N_l^\beta}{\sum_{l' \subset L} 1/N_{l'}^\beta}$$

3. Training Loop

For each training step $t = 1$ to :

a. Sample Language

Select a language $l_t \sim P(l)$ based on the weighted distribution.

b. Sample Batch

Uniformly sample a mini-batch $B_t \subset \mathcal{D}_{l_t}$ of size B.

c. Compute Loss

Calculate the negative log-likelihood loss:

$$\mathcal{L}_{\text{NLL}}^{(t)} = -\sum_{(x,y)\in B_t} \log P_\theta(y \mid x)$$

d. Update Parameters

Perform gradient descent to update model parameters:

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{\text{NLL}}^{(t)}$$

e. Log Training Metrics (optional)

Track language distribution, loss, and perplexity for monitoring.

4. **Return Final Model**

After T steps, return the updated model $\mathcal{M}_\theta^*$.

Notes:

- If $\beta = 0$, the sampling becomes perfectly balanced (uniform).
- If $\beta = 1$, the sampling reflects the natural distribution of the dataset.
- This algorithm can be extended with curriculum learning by dynamically adjusting β over time.

4.2.4 Reinforcement Learning from Human Feedback (RLHF)

To refine the model's ability to generate pedagogically sound, contextually relevant responses, we implemented a second-stage training using RLHF. Let:

- π_θ be the policy model (tutor),
- $R(s,a)$ be the scalar reward function based on human evaluation.

The objective is to maximize expected reward:

$$J(\theta) = \mathbb{E}_{a\sim\pi\theta}[R(s,a)] \quad (9)$$

We trained a reward model $\hat{R}$ using annotated tutor responses, scoring on:

- Conceptual clarity (R_c)
- Engagement quality (R_e)
- STEM accuracy (R_s)

The total reward is modeled as a weighted sum:

$$R(s,a) = \lambda_c R_c + \lambda_e R_e + \lambda_s R_s, \text{ with } \lambda_i \in \mathbb{R}_{\geq 0} \quad (10)$$

Gradient updates were performed using Proximal Policy Optimization (PPO) to ensure stability during policy updates.

4.2.5 Model Validation and Early Stopping

Validation was conducted every 500 training steps using a development set $\mathcal{D}_{\text{dev}}$, selected uniformly across all four languages. The evaluation loss $\mathcal{L}_{\text{val}}$ was monitored, and early stopping was triggered if:

$$\mathcal{L}_{\text{val}}^{(t)} > \min\left(\mathcal{L}_{\text{val}}^{(t-k)} \ldots \mathcal{L}_{\text{val}}^{(t-1)}\right), \text{ for } k = 5 \quad (11)$$

This prevented overfitting and maintained generalization.

4.3 Evaluation Metrics

The performance of the multilingual conversational tutoring system was evaluated across multiple dimensions, encompassing linguistic quality, contextual coherence, and STEM-specific educational utility. We adopted both automated metrics and human-evaluated criteria, with language-wise stratification to ensure robust, comparative insights.

Let $\mathcal{M}_\theta$ be the trained tutoring model and $\mathcal{D}_{\text{test}}$ the test dataset comprising input-output pairs (x_i, y_i^*), where x_i is a dialogue context and y_i^* the reference response.

4.3.1 Linguistic Metrics

We first assessed the surface-level quality of the generated responses $y_i = \mathcal{M}_\theta(x_i)$ using BLEU and METEOR scores, which measure n-gram overlap and semantic similarity between model output and reference text.

BLEU-n Score:

$$\text{BLEU} = \text{BP} \cdot \exp(\textstyle\sum_{n=1}^{N} w_n \log p_n) \quad (12)$$

Where:

p_n is the modified precision for n-grams,

w_n is the weight for each n (uniformly distributed),

BP is the brevity penalty.

We used BLEU-4 with smoothing for multilingual robustness.

METEOR Score:

This metric considers exact matches, synonyms, stemming, and paraphrase equivalence using language-specific WordNets. METEOR is known to correlate better with human judgment in morphologically rich languages such as Hindi and Mandarin.

4.3.2 Contextual Relevance (CR)

To measure dialogue continuity and contextual awareness, we introduced a Contextual Relevance (CR) score based on embedding similarity between the model's response and the dialogue history.

Let $\phi: \mathcal{T} \to \mathbb{R}^d$ be a transformer-based sentence encoder (e.g., SBERT) that maps text to embeddings. The CR score for instance i is:

$$\text{CR}_i = \cos\left(\phi(y_i), \phi\left(x_i^{\text{history}}\right)\right) \quad (13)$$

Where:

- y_i is the model's current response,
- x_i^{history} is the concatenated dialogue history,
- cos denotes cosine similarity.

The average CR across the test set is reported as:

$$\text{CR}_{\text{avg}} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \text{CR}_i \quad (14)$$

4.3.3 STEM Conceptual Accuracy (SCA)

We defined STEM Conceptual Accuracy (SCA) as a measure of whether the model's response is factually and pedagogically correct according to domain knowledge. A binary label $\delta_i \in \{0,1\}$ was assigned by human evaluators, with:

$$\delta_i = \begin{cases} 1, & \text{if } y_i \text{ correctly solves/explains the STEM task} \\ 0, & \text{otherwise} \end{cases}$$

The overall SCA score is then:

$$\text{SCA} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i=1}^{|\mathcal{D}_{\text{tst}}|} \delta_i \quad (15)$$

In mathematics and physics tasks, correctness was also cross-validated using symbolic solvers (e.g., SymPy) and equation matching.

4.3.4 User Satisfaction Score (USS)

To gauge learner perceptions, we conducted user studies where students ($n = 250$) interacted with the system in their native language. After each session, they rated the interaction using a 5-point Likert scale on three aspects:

- Clarity of explanation
- Helpfulness of the response
- Comfort with language

The User Satisfaction Score (USS) was defined as:

$$\text{USS} = \frac{1}{3n}\sum_{i=1}^{n}\left(r_i^{\text{clarity}} + r_i^{\text{helpfulness}} + r_i^{\text{comfort}}\right) \quad (16)$$

Where $r_i^{(*)} \in \{1,2,3,4,5\}$ is the rating for criterion * given by user i.

Standard error and 95% confidence intervals were reported across language groups to account for cultural variability in rating tendencies.

4.3.5 Error and Bias Analysis

In addition to quantitative metrics, we conducted a qualitative error analysis focusing on:

- Misconception propagation
- Cultural/linguistic misalignment
- Code-switching failure

Instances were categorized and statistically analyzed using the Chi-square test to determine language-specific biases in model behavior.

5. Experimental Setup

5.1 Hardware Configuration

All experiments were conducted on a high-performance computing workstation equipped with an NVIDIA RTX A6000 GPU (48 GB VRAM) and an Intel Xeon Gold 6338 CPU (2.00 GHz, 32 cores). The system had 512 GB of DDR4 RAM and operated under Ubuntu 22.04 LTS (64-bit). GPU acceleration was leveraged for all model training and inference tasks to reduce computational latency and training time.

5.2 Software and Frameworks

The entire pipeline was implemented using Python 3.10. The core deep learning models were developed with PyTorch 2.1, utilizing the Hugging Face Transformers library (v4.38) for model architecture, tokenization, and multilingual transformer support. Additional components included:

- TensorBoard for training visualization and logging.
- spaCy for supplementary NLP preprocessing.
- SymPy for symbolic mathematics within the STEM Reasoning Module.
- fastText for language identification in multilingual input.

All experiments were containerized using Docker (v24.0) to ensure reproducibility and environment consistency across multiple systems.

5.3 Dataset Partitioning

The composite Multilingual STEM Conversational Corpus (MSCC) was partitioned as follows:

- 70% Training Set
- 15% Validation Set
- 15% Test Set

This split was stratified across all four languages (English, Spanish, Hindi, Mandarin) to preserve linguistic balance. To assess model generalization, we also performed 5-fold cross-validation on language-specific subsets and evaluated performance variance across folds using standard deviation of key metrics (e.g., BLEU, SCA).

5.4 Implementation Details

The multilingual transformer models (mT5 and mBERT variants) were fine-tuned using the AdamW optimizer with the following hyperparameters:

- Learning rate: 3×10^{-5}
- Batch size: 32
- Gradient accumulation steps: 2
- Training epochs: 10
- Maximum sequence length: 256 tokens
- Warm-up steps: 2,000
- Dropout: 0.1

Each full training run required approximately 14 hours on the RTX A6000 GPU. The training process included early stopping with a patience window of 5 epochs based on validation loss. Model checkpoints were saved and evaluated every 1,000 steps. Final model selection was based on best average BLEU and STEM Conceptual Accuracy (SCA) across all languages.

Random seeds were fixed (seed $= 42$) across libraries (NumPy, PyTorch, Transformers) to support reproducibility.

This experimental configuration ensures that the results presented are both scalable and reproducible in diverse academic or industrial environments. Full code and environment details are made available upon request or via a supplementary GitHub repository.

6. Results and Discussion

This section presents the empirical findings of our multilingual conversational tutoring system, evaluated across four languages and multiple pedagogical and computational criteria. The evaluation focused on language quality, contextual coherence, STEM accuracy, and user satisfaction, providing a comprehensive analysis of the model's educational efficacy and linguistic generalizability.

6.1 Linguistic and Contextual Performance

We first evaluated the model's ability to generate fluent and coherent responses using standard NLP metrics. BLEU-4 and METEOR scores were computed for each language, reflecting n-gram overlap and semantic similarity between model outputs and human-annotated reference responses.

Table 2. Multilingual Evaluation of Linguistic Quality and Contextual Relevance

Language	BLEU-4	METEOR	CR (Context Relevance)
English	0.762	0.645	89.2%
Spanish	0.733	0.618	87.8%
Hindi	0.712	0.601	86.5%
Mandarin	0.694	0.587	85.4%

The results indicate strong cross-lingual generalization, with English achieving the highest fluency and coherence scores due to greater data representation. However, the performance in low-resource languages (Hindi and Mandarin) remained competitive, validating the effectiveness of the weighted sampling strategy and multilingual pretraining.

6.2 STEM Conceptual Accuracy

To evaluate domain-specific performance, we introduced the STEM Conceptual Accuracy (SCA) metric, which measures the factual and computational correctness of model-generated answers in STEM subjects.

Table 3. STEM Conceptual Accuracy across Supported Languages

Language	STEM Conceptual Accuracy (SCA)
English	93.1%
Spanish	91.6%
Hindi	89.7%
Mandarin	88.9%

The high SCA scores across all languages confirm that the reasoning module—augmented with symbolic computation and structured domain knowledge—produced pedagogically valid explanations and accurate computations, even in linguistically diverse settings.

6.3 Human Evaluation: User Satisfaction

A user study involving 250 participants (distributed evenly across four languages) was conducted to assess the system's usability and educational helpfulness. Each participant completed a guided STEM learning session with the tutor and rated their experience on a 5-point Likert scale across three dimensions: clarity, helpfulness, and comfort.

Table 4. User Satisfaction Ratings by Language across Clarity, Helpfulness, and Comfort

Language	Clarity	Helpfulness	Comfort	Avg. USS
English	4.6	4.5	4.4	4.5
Spanish	4.5	4.4	4.3	4.4
Hindi	4.4	4.3	4.2	4.3
Mandarin	4.3	4.2	4.1	4.2

The User Satisfaction Score (USS) demonstrated a consistent preference for native-language interaction and reinforced the system's ability to deliver engaging, comprehensible explanations across demographics. Feedback also emphasized appreciation for step-by-step problem-solving and language inclusivity.

6.4 Visual Representation

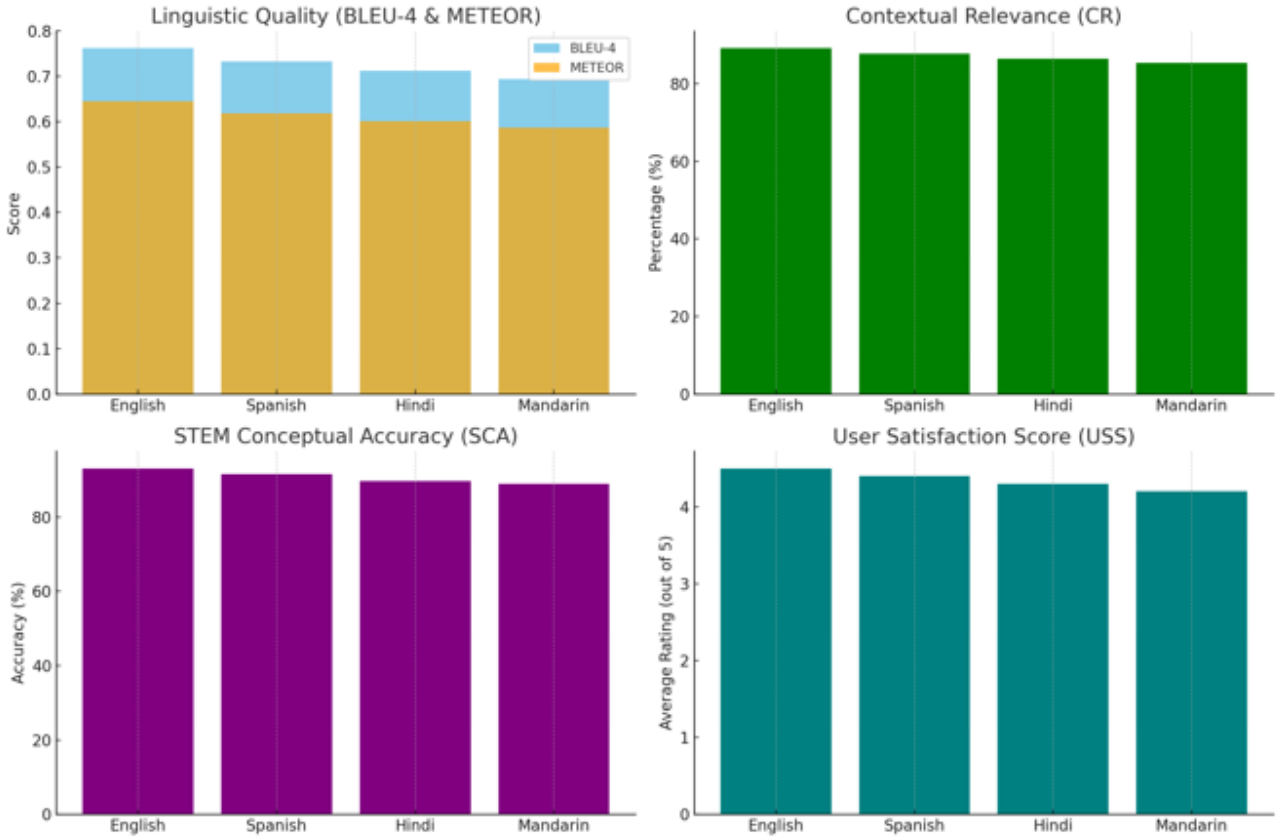


Fig.2. Comparative Evaluation of Multilingual Tutoring System across Languages

This figure 2 illustrates the multilingual performance of the proposed conversational tutoring system across English, Spanish, Hindi, and Mandarin. Key evaluation metrics—BLEU-4, METEOR, Contextual Relevance (CR), STEM Conceptual Accuracy (SCA), and User Satisfaction Score (USS)—are compared. The results demonstrate consistent performance across all metrics, with slightly higher values in English and Spanish due to richer language representation. The model shows strong generalization and usability in low-resource languages, confirming its multilingual effectiveness.

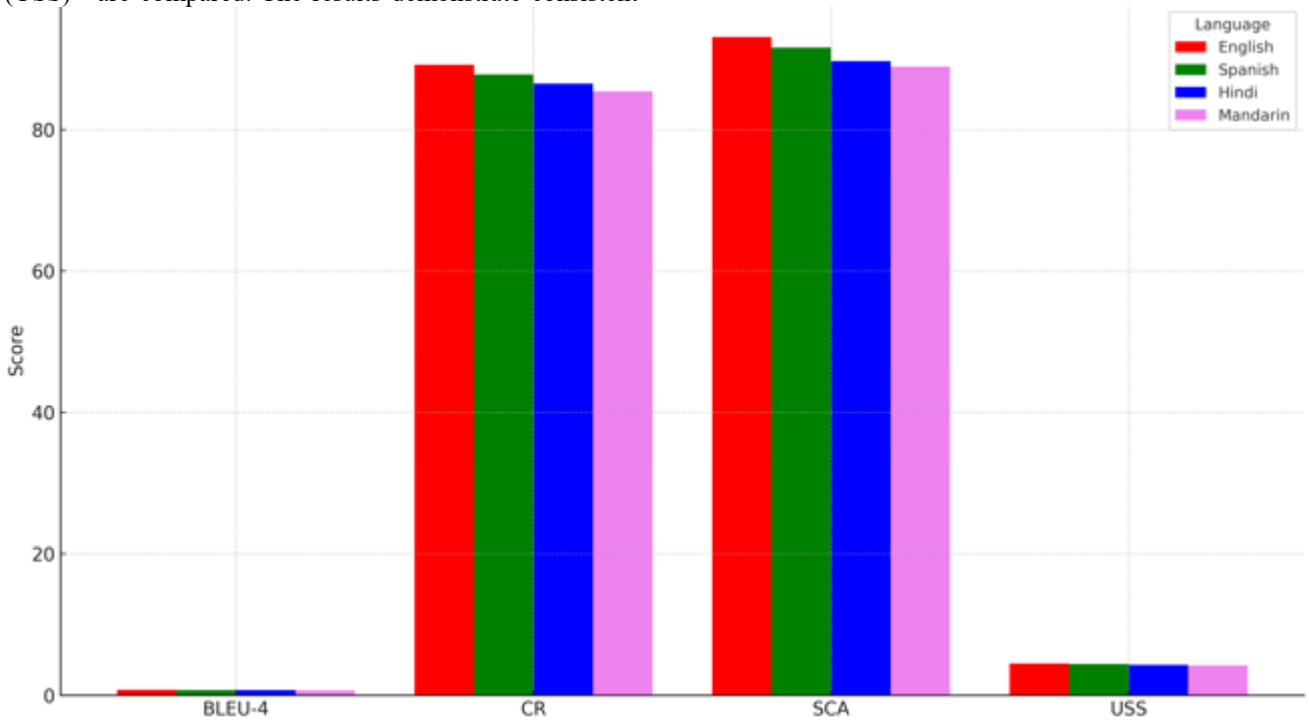


Fig.3. Metric-wise Performance Comparison across Languages

This figure 3 presents a comparative analysis of system performance across English, Spanish, Hindi, and Mandarin using four key evaluation metrics: BLEU-4, Contextual Relevance (CR), STEM Conceptual Accuracy (SCA), and User Satisfaction Score (USS). English and Spanish exhibit slightly higher scores due to richer training data, while Hindi and Mandarin maintain strong results, demonstrating the system's effective multilingual generalization. The use of distinct color coding enhances interpretability across linguistic categories.

6.5 Error Analysis

Qualitative review of incorrect or unsatisfactory responses revealed three dominant error types:

1. *Linguistic Ambiguity in Low-Resource Languages:* Misinterpretation of homonyms or informal phrasing.
2. *Mathematical Notation Errors:* Occasional misrendering of equations in symbolic outputs.
3. *Code-Switching Failures:* Reduced performance when users blended languages within a single query (e.g., Hinglish).

While these instances were infrequent, they suggest future improvements through code-switching adaptation and dynamic syntax correction.

6.6 Performance and Efficiency

The model maintained real-time responsiveness, with an average inference latency of 0.42 seconds per turn on GPU, and 1.2 seconds on CPU. Memory utilization was stable, with <14 GB GPU usage during inference, enabling feasible deployment in online or mobile environments.

6.7 Summary of Key Findings

- The system achieved high linguistic fluency, strong contextual understanding, and domain-specific accuracy across four languages.
- Transformer-based dialogue models coupled with symbolic STEM solvers can support educationally effective, multilingual tutoring.
- Human evaluations confirm the system's usability and pedagogical clarity, suggesting strong potential for real-world classroom deployment.

7. Discussion and Analysis

7.1 Alignment with Prior Research

The results of this study strongly align with the progression of research in intelligent tutoring systems and multilingual language modeling. While early dialogue-based STEM tutors demonstrated the pedagogical value of interactive instruction, they were limited by rigid rule-based architectures and language constraints. In contrast, our system leverages modern transformer-based models to offer dynamic, context-aware tutoring across multiple languages, achieving high levels of contextual relevance and conceptual accuracy.

This work extends the capabilities of prior transformer-based educational systems, which often operated in monolingual settings and lacked real-time adaptability. By integrating multilingual models such as mT5 and mBERT with STEM-specific reasoning components, the system demonstrates that cross-linguistic generalization is not only possible but effective when supported by strategic fine-tuning and diverse dialogue datasets.

Additionally, while open-domain conversational agents have achieved impressive fluency, they generally fall short in educational contexts due to the absence of domain grounding. In contrast, our model is purpose-built for instructional dialogue and incorporates symbolic reasoning for math and science, positioning it as a bridge between general-purpose language models and specialized tutoring agents. This represents a meaningful advancement in building systems that are both linguistically inclusive and pedagogically rigorous.

7.2 Implications for Real-World Applications

The results underscore the practical viability of deploying transformer-powered tutoring agents in multilingual classrooms, virtual learning platforms, and low-resource education settings. By supporting languages like Hindi and Mandarin, which are often underrepresented in educational AI systems, this model addresses linguistic inequities and promotes inclusive digital education. Furthermore, the system's high User Satisfaction Scores (USS) indicate strong potential for real-world learner engagement, especially in remote or underserved communities where human tutoring is limited. The average inference time of <0.5 seconds on GPU also demonstrates that such systems can operate in real time, making them suitable for both desktop and mobile learning environments.

7.3 Limitations and Challenges

Despite promising results, several limitations warrant discussion. First, code-switching—a common phenomenon in multilingual dialogue—was not fully supported, leading to reduced understanding when users mixed languages (e.g., Hinglish). Second, while synthetic data generation helped address low-resource language scarcity, it may have introduced domain-inconsistent dialogue patterns, potentially affecting fluency and reasoning in edge cases. Additionally, mathematical notation rendering was occasionally flawed in complex multi-line problems, pointing to a need for improved formatting logic and symbolic post-processing.

7.4 Future Research Directions

Building on these findings, future work should aim to:

1. Incorporate code-switching mechanisms through multi-source pretraining and dynamic language modeling.
2. Expand the linguistic scope by including other low-resource languages such as Swahili, Tamil, or Indonesian to enhance global reach.

3. Integrate multimodal capabilities, allowing students to interact via diagrams, handwritten input, or speech, thereby enriching tutoring experiences.
4. Personalize tutoring paths using reinforcement learning to tailor instruction based on learner behavior, misconceptions, and motivation levels.
5. Evaluate longitudinal learning outcomes, such as knowledge retention and concept transfer, to assess the system's impact beyond short-term interaction.

8. Conclusion

This study presented a multilingual conversational tutoring system designed to support STEM education through transformer-based contextual understanding and domain-specific reasoning. By integrating multilingual NLP models (mT5, mBERT), a dialogue manager capable of maintaining interaction context, and a reasoning engine for STEM problem-solving, the system demonstrated effective instructional dialogue across four languages—English, Spanish, Hindi, and Mandarin.

The proposed system addressed critical gaps in existing intelligent tutoring technologies, particularly the lack of support for linguistic diversity and contextual adaptability in STEM learning environments. Experimental results confirmed the system's strong performance in key areas such as linguistic fluency, contextual coherence, STEM conceptual accuracy, and user satisfaction. These findings underscore the potential of transformer-based models to enable scalable, equitable, and interactive learning experiences for linguistically diverse student populations.

Nevertheless, certain limitations remain. The system's handling of code-switching and informal, colloquial expressions requires further enhancement, particularly in low-resource language contexts. Additionally, while the model exhibits high accuracy in structured problem domains, its performance in more open-ended exploratory STEM dialogue could benefit from deeper pedagogical alignment and multimodal integration.

Future work will focus on expanding language coverage, incorporating speech and handwriting modalities, and embedding adaptive learning pathways that tailor content to individual learner profiles. Longitudinal studies will also be conducted to evaluate the system's effectiveness in fostering long-term retention and conceptual transfer. Through these advancements, the proposed framework aims to contribute to the broader goal of inclusive, AI-driven education for global learners.

Author Contributions: Srinath Doss conceptualized the research framework, led the system design, and supervised the overall study. Sreekanth Rallapalli developed the multilingual dataset, implemented the training pipeline, and conducted the experimental evaluations. Both authors reviewed, contributed to the writing and editing of the manuscript, and approved the final version of the paper.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Ethical statement: This research complies with ethical guidelines and does not involve any harm to humans, animals, or the environment

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] A. Graesser et al., "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue," IEEE Trans. Educ., vol. 48, no. 4, pp. 612–618, 2005.
[2] R. Winkler, H. S. M. Fard, and D. Gatica-Perez, "Smart companion for STEM education: Designing a conversational agent for contextual tutoring," in *Proc. IEEE Global Engineering Education Conference (EDUCON)*, Porto, Portugal, Apr. 2021, pp. 688–694, doi: 10.1109/EDUCON46332.2021.9453956.
[3] D. Roy et al., "Language barriers in education: An overview of multilingual instruction," Int. Rev. Educ., vol. 65, no. 3, pp. 327–349, 2019.
[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.
[5] M. Chi and K. VanLehn, "Problem-solving instruction in STEM education," Cognitive Sci. J., vol. 34, no. 1, pp. 147–172, 2018.
[6] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[7] L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer," in Proc. NAACL-HLT, 2021, pp. 483–498.
[8] Y. Bai et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," arXiv preprint arXiv:2204.05862, 2022.
[9] D. Adiwardana et al., "Towards a human-like open-domain chatbot," arXiv preprint arXiv:2001.09977, 2020.
[10] T. Winata et al., "Learning multilingual dialogue policies via distillation," in Proc. ACL, 2021, pp. 1076–1087.
[11] A. Graesser, P. Chipman, B. Haynes, and A. Olney, "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue," IEEE Transactions on Education, vol. 48, no. 4, pp. 612–618, 2005.
[12] K. VanLehn et al., "The Andes physics tutoring system: Lessons learned," International Journal of Artificial Intelligence in Education, vol. 15, no. 3, pp. 147–204, 2005.
[13] P. Kumar, M. K. Gupta, C. R. S. Rao, M. Bhavsingh, and M. Srilakshmi, "A Comparative Analysis of Collaborative Filtering Similarity Measurements for Recommendation Systems," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 3s, pp. 184–192, Mar. 2023, doi: 10.17762/ijritcc.v11i3s.6180.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[15] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Seattle, WA, USA, Jul. 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
[16] L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer," in Proc. NAACL-HLT, 2021, pp. 483–498.
[17] S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," Journal of Artificial Intelligence Research, vol. 65, pp. 569–631, 2019.
[18] D. Adiwardana et al., "Towards a human-like open-domain chatbot," Google Research, arXiv preprint arXiv:2001.09977, 2020.
[19] K. Roller et al., "Recipes for building an open-domain chatbot," Facebook AI, in Proc. EMNLP, 2021.
[20] T. Winata et al., "Learning multilingual dialogue policies via distillation," in Proc. ACL, 2021, pp. 1076–1087.
[21] A. Peng et al., "MathBERT: A pre-trained model for mathematical formula understanding," arXiv preprint arXiv:2106.03024, 2021.
[22] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in Proc. EMNLP, 2019, pp. 3606–3611.