CTU-Tutor: Contextual and User-Aware Language Understanding for Personalized Tutoring with Long-Form Text Generation

Akanksha Bisht, Oshin Sharma
Department of Computer Science and Engineering-Faculty of Engineering and Technology,
SRM Institute of Science and Technology NCR Campus, Delhi-NCR Campus,
Delhi-Meerut Road, Modinagar, Ghaziabad, U.P, India

***Abstract*—This study introduces CTU-Tutor, a contextual and user-aware intelligent tutoring system for generating personalized long-form text as learning content. The system combines BERT-based learner embeddings, K-means for proficiency clustering, Transformer-XL for long-context modelling, BiLSTM-CRF for concept extraction, and LDA-based knowledge graph construction. A retrieval-augmented Transformer-XL model produces personalized content, which is supplemented with explainable AI to make decisions transparently using SHAP and LIME. Experimental analysis shows that CTU-Tutor is better than state-of-the-art models, such as ExPerT, REST-PG, GSPT-CVAE, LONGLaMP, and Transformer_QA, in various metrics. The proposed framework has 0.98 accuracy, 0.98 F1-score, 0.97 MCC, and 0.98 precision, sensitivity, and specificity, and minimized false negatives (0.013) and false positives (0.017). For text generation, BLEU, ROUGE, and METEOR scores were approximately 0.98, indicating high overlap with the corresponding human-authored reference texts. These results prove that the synergistic implementation of learner profiling, long-context modelling, structured knowledge representation, and explainable AI delivers more precise, reliable, and pedagogically transparent personalized tutoring.**

***Keywords*—Contextual and user-aware tutoring system; bidirectional encoder representations of transformers; transformer-XL; bidirectional long short-term memory with conditional random fields; explainable AI; latent dirichlet allocation**

I. Introduction

Development of intelligent tutoring systems (ITS) that facilitate individualized learning experiences has been sped up by the quick development of artificial intelligence (AI) and natural language processing (NLP) [1]. AI-driven tutors provide opportunities for individualized guidance, scalable learning support, and ongoing assessment, in contrast to traditional classroom instruction, where teaching is frequently standardized and teacher attention is distributed among numerous students [2]. Adaptive systems are now essential to contemporary e-learning ecosystems due to the growing availability of digital educational content and learner interaction data [3]. This allows for content delivery that is responsive to learner profiles, progress, and preferences.

The capacity of current AI tutoring systems to produce lengthy, pedagogically structured, and contextually tailored content is still restricted, despite tremendous advancements [4]. The majority of the most advanced models, especially Transformer-based large language models (LLMs), concentrate on producing brief responses that frequently lack pedagogical depth and conformity to the proficiency levels of the learners [5,6]. Even though hybrid and retrieval-based tutoring systems have strengthened their factual foundation, they are still unable to meet the complex needs of educational personalization, like adjusting to a learner's preferred learning pathway, cognitive state, or past knowledge [7,8]. In addition, deep learning models' opaque decision-making creates obstacles to adoption and trust in authentic learning environments, where both educators and students expect clear and understandable systems [9,10].

The primary goals of the research are to design and create a personalized and context-aware tutoring system that produces long-form learning content that is specific to individual learners. To be more precise, the research develops learner embeddings and proficiency bands to use in adaptive instruction, deploys content understanding and knowledge extraction with the help of advanced NLP models, generates pedagogically coherent long-form explanations, and adds an explainable AI (XAI) layer to present transparent reasons and confidence scores, which contribute to the effectiveness of learning and increase the level of trust.

This study presents the CTU-Tutor, a novel framework that combines contextual knowledge extraction, learner profiling, content understanding, and explainable generation of long-form instructional material to address these issues. In contrast to previous methods, CTU-Tutor uses a customized Transformer-XL architecture in conjunction with knowledge graph-driven content structuring, proficiency-based clustering, and user-aware embeddings. The framework also integrates an XAI layer through SHAP and LIME, which offers both global and local interpretability, and uses curriculum learning with human feedback to guarantee pedagogical coherence. What sets CTU-Tutor apart from other AI tutors is its comprehensive fusion of explainability, context, and personalization.

The main contributions of this study are as follows:

- A learner profiling mechanism is developed by integrating BERT-based learner representations with K-Means proficiency clustering to identify learner proficiency bands and incorporate learner-specific

characteristics into personalized instructional content generation.

- A unified content-generation mechanism is developed by connecting learner-aware representations with contextual language modelling, concept extraction, and structured educational knowledge representation to support contextualized long-form instructional content generation.
- An explainability layer is integrated into the tutoring pipeline to provide interpretable explanations of learner-related predictions and generated instructional decisions, thereby improving transparency and supporting the pedagogical acceptability of the system.
- The above components are integrated into a unified tutoring pipeline that combines learner proficiency modelling, structured educational knowledge, contextual generation, retrieval-based grounding, and explainability.

The remainder of this article is structured as follows. Section II reviews related works in personalized tutoring, NLP-based educational systems, and explainable AI. Section III presents the proposed CTU-Tutor framework. Section IV presents results and comparative analysis. Finally, Section V concludes the study and summarizes key contributions.

II. Literature Survey

Lin et al. [11] introduced the EPER (Enhanced Personalization for Explainable Recommendation) model that used Transformer models with reviews, user ratings, feature words, and item titles to produce personalized explanations. This was to improve the quality and personalization of recommendations as well as overcome the weaknesses of the previous RNN-based approaches. Nevertheless, it had constraints in dealing with the sparsity of data and in adapting effectively to large-scale real-life recommendation problems.

Sapin et al. [12] proposed a transformer-based system that uses large language models (LLMs), specifically GPT-4 with prompt engineering, to recognize dialogue acts in team communication. This was done to improve adaptive training systems by increasing the accuracy of classification with a small amount of data. Nevertheless, it had drawbacks associated with the use of small datasets and potential limitations.

Vidakovic et al. [13] presented fine-tuned multilingual transformer models with data augmentation methods to examine Serbian student feedback in an Intelligent Tutoring System. This was to improve software performance and learning plans by means of automated feedback analysis. Nonetheless, the research had a limitation regarding language-specific generalizability and the reliance on annotated corpora to ensure the same performance.

Jia and Lee [14] proposed five methods of dealing with long-text sequences in organizational NLP studies, such as TF-IDF-RF, Longformer, GPT-4o, truncation with averaged scores, and a construct-relevant text-selection method. This was done to assess their predictive validity and psychometric properties. Nevertheless, the research had its limitations because no specific method was always better than others, which is why it is necessary to refine the research and make it context-specific.

Li et al. [15] proposed a large-scale benchmark called PERSONACONVBENCH that measures personalized reasoning and generation on multi-turn conversations. It was performed with the aim of measuring the effects of personalization in various areas systematically. The benchmark showed large performance improvements, but it was still constrained by Reddit-specific data and the high cost of evaluation.

Sharma et al. [16] surveyed how Large Language Models are transforming personalised learning as it provides learning paths according to the user's needs, styles, and preferences. The methods use user profiles, past interactions, and personalized goals with the help of prompting, fine-tuning adapters, and preference alignment. Despite the reported significant improvements, there were still challenges in terms of data privacy, scalability, and robustness when the user data is sparse or noisy.

Zhang et al. [17] have discussed the lack of cohesion in personalised LLM research by proposing a systematic taxonomy that consolidated personalized text generation and personalization-based downstream applications. The method formalized the underlying ideas, granularity of personalization, procedures, datasets, and assessment procedures to integrate various research strands. Although important for standardization and clarity, the methodology had weaknesses in conceptual abstraction and a lack of empirical validation on large-scale deployments.

Vinod et al. [18] proposed INVISIBLEINK, a differentially private text generation model that safely injected sensitive information into the text generated by an LLM. The method was based on a token sampling framework that modelled it as an exponential sampling mechanism. Although the efficiency gains were strong, limitations such as the complexity of added systems and the dependence on the careful separation of the public and the private signals were also present.

Baradari et al. [19] suggested NeuroChat, a neuroadaptive AI tutoring system combining real-time engagement monitoring using EEG with generative language models. The method used a closed-loop system to dynamically modify the content complexity, pacing, and style of response to the cognitive state of the learners. Although the results of pilots indicated better cognitive and subjective engagement, the small-scale assessment and lack of immediate improvement in the learning outcomes were considered limitations.

Wu et al. [20] presented the LongGenBench framework to evaluate the long-form text generation ability of LLMs rather than simple text generation. It generates long outputs up to 16k- 32k tokens while following multistep instructions, such as inserting events or constraints at specific positions in the text. It checks whether the framework only focuses on generating text of long length or adheres to the provided instructions across the generation process.

Rudik and Onyshchuk [21] studied the application of AI and large language models to English as a Second Language (ESP) education, including methods that used NLP and generative models to generate customized and profession-specific learning resources. The strategy was to increase the level of learner engagement, contextual relevance, and focused language acquisition through human-AI interaction. Although pedagogically important, such limitations as risk of bias, factual errors, and difficulties in the responsible incorporation of AI into the teaching practice of instructors were identified.

Safeni et al. [22] developed a game-theoretic model of controllable text generation that controlled the language model outputs by adaptive prompt interventions. The method used Nash equilibrium principles and feedback to optimize dynamically relevant diversity and coherence in the text generated. Despite its effectiveness and adaptability, the method was constrained by the complexity of the stimulation in the prompt and the dependence on the well-thought-out choice of the strategy. Table I presents existing models along with findings and associated limitations.

TABLE I. SUMMARY OF RELEVANT MODELS, FINDINGS, AND LIMITATIONS

Model	Findings	Disadvantages
ExPerT [23]	Improved alignment with human judgments by 7.2% while providing highly interpretable, fine-grained explanations for personalized text assessment.	Additional computational cost and potential bias in the evaluation process.
REST-PG [24]	Improved personalized long-form text generation by enabling LLMs to explicitly reason over user preferences	Increased training complexity and computational overhead
GSPT-CVAE [25]	Improved controllable long-text generation by learning rich graph-structured textual relationships and effective potential representations	Increased architectural complexity and computational overhead
LONGLaMP [26]	User-specific personalization across diverse personalized long-text generation tasks	Required substantial personalized data
Transformer_QA [27]	Accurate interpretation of student queries and enhanced engagement, motivation, and learning outcomes	Predefined rules, limited scalability, adaptability to diverse subjects

A. Problem Statement

Despite recent developments in AI-powered tutoring systems, the majority of current methods find it difficult to provide long-form educational content that is understandable, contextually aware, and personalized for each student. Either generic short-form responses are provided by current models, learner profiling and interpretability are not incorporated into the content generation process, or mechanisms for deep content understanding and knowledge structuring are absent. Because diverse learners need adaptive instruction that is in line with their learning context, pedagogical transparency, and proficiency levels, this leads to a substantial disconnect between automated tutoring systems and their real needs. For efficient and reliable individualized tutoring, a framework that combines user profiling, contextual knowledge extraction, personalized content creation, and explainable AI into a single pipeline must be created.

III. PROPOSED FRAMEWORK: CTU-TUTOR

Fig. 1 illustrates the general architecture of the proposed CTU-Tutor framework, which is organized into four large interconnected layers, each of which is essential in facilitating adaptive, personalized, and explainable tutoring. User Profiling and Context Understanding is the first layer, which provides the basis of the learner representation by incorporating both demographic and interactional information to create a holistic picture of the learner.

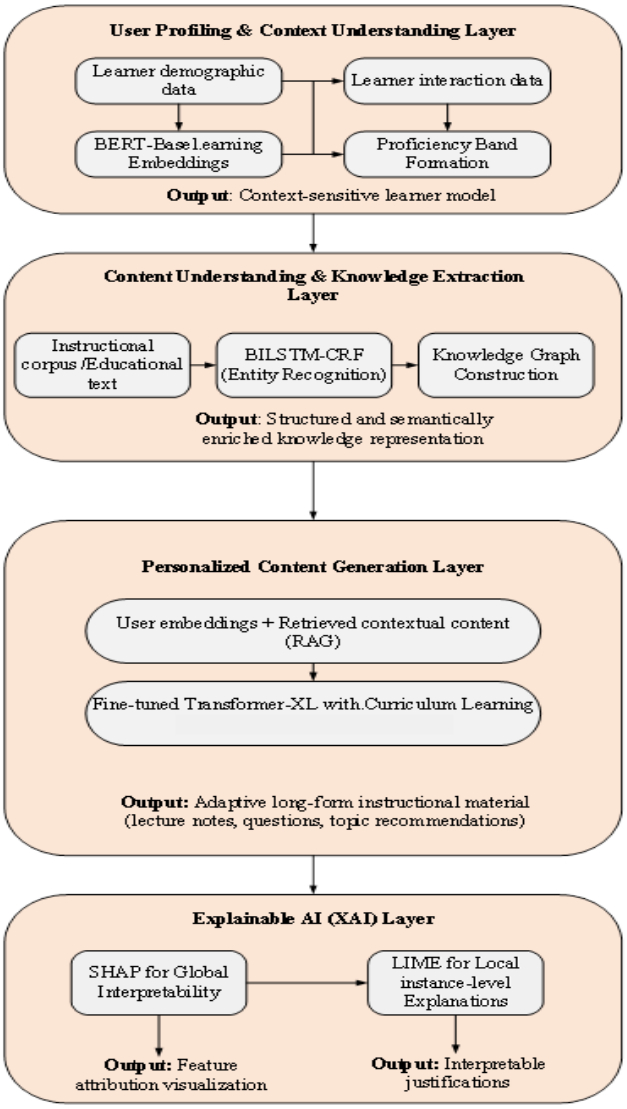


Fig. 1. Architectural design of the proposed CTU-Tutor framework.

The information is embedded using a BERT-based embedding model that is fine-tuned on educational metadata to transform heterogeneous information about learners into semantic numerical representations. These embeddings are then clustered using the K-Means algorithm to form proficiency bands, enabling the system to classify learners according to their skill levels and behavioural traits. The output of this layer is a context-sensitive learner model, which is used as the foundation of personalization in the framework. Content Understanding and Knowledge Extraction is the second layer,

which is concerned with the conversion of raw instructional materials into structured pedagogical knowledge. It starts with Transformer-XL, which classifies sentences at the sentence level to take into account long-range relationships in text, and then BiLSTM-CRF, which identifies the main educational concepts and relationships with the help of precise entity recognition and sequence labelling. The extracted entities are then semantically clustered using LDA to create a topic-based representation, which drives a Knowledge Graph Construction module that generates a structured and semantically enriched knowledge representation of the educational corpus. Personalized Content Generation is the third layer, which combines the information of user profiles and content structures to create adaptive instructional content. It integrates user embeddings from the first layer with contextual outputs retrieved via RAG, and uses a fine-tuned Transformer-XL trained under curriculum learning principles and human feedback alignment to produce long-form, pedagogically coherent outputs such as lecture notes, topic recommendations, and question-answer sets. Finally, the XAI Layer enhances transparency and trust by incorporating SHAP for global interpretability and LIME for local, instance-based explanations. This layer outputs feature attribution visualizations and interpretable justifications, helping educators and learners understand the rationale behind each recommendation and generated content. Overall, the proposed CTU-Tutor integrates user modelling, deep content understanding, personalized generation, and explainable AI into a unified, intelligent tutoring pipeline that ensures contextual relevance, pedagogical coherence, and interpretability in every stage of personalized learning.

A. User Profiling and Context Understanding

During this stage, the demographic data as well as the interaction data are gathered to ensure that a holistic representation of the learner is achieved. This is critical in modelling not only the knowledge state of the learner but also their behaviour patterns, which have a direct impact on the adaptation of instructional content. To represent the information about the learners formally, we define the demographic data as a feature vector:

$$D_i = \{d_{i1}, d_{i2}, \dots, d_{im}\}, i = 1,2, \dots, N \quad (1)$$

where, D_i denotes the demographic attributes of the learner i across m features. Similarly, the interaction data is represented as:

$$I_i = \{i_{i1}, i_{i2}, \dots, i_{im}\}, i = 1,2, \dots, N \quad (2)$$

where, I_i captures the sequence of interaction features for the learner i. The input feature set of the user embedding is the combination of both demographic and interaction features:

$$X_i = D_i \cup I_i \quad (3)$$

A BERT-base model [28] that is fine-tuned on learner metadata is used to convert this heterogeneous data into meaningful numerical representations. Let f_{BERT} represent the embedding function; the learner embedding for the user i is obtained as:

$$E_i = f_{BERT}(X_i) \in R_k \quad (4)$$

where, E_i is the embedding vector of dimension k, encapsulating the semantic and contextual representation of the learner's profile and learning behavior. These embeddings enable the framework to process learner profiles not as discrete categorical features but as high-dimensional and context-rich representations, which are used in downstream tasks like clustering and personalization.

Once embeddings are generated, the next step is to categorize learners into proficiency bands using the K-Means clustering algorithm [29]. The clustering objective is to partition the embedding space into K clusters $\{C_1, C_2, \dots, C_K\}$ such that intra-cluster similarity is maximized while inter-cluster similarity is minimized. Mathematically, the optimization function of K-Means is given by:

$$\underset{c}{argmin} \sum_{j=1}^{k} \sum_{E_i \epsilon c_j} \|E_i - \mu_j\|^2 \quad (5)$$

where, μ_j denotes the centroid of the cluster C_j. To avoid circularity between the learner representation and proficiency labels, we separate historical input variables from target variables. Let $X_i^{(t)}$ denote the historical features available for the learner i at time t, constructed as in Eq. (3) from demographic and interaction information up to time t. The learner representation is computed as:

$$E_i^{(t)} = f_{\text{BERT}}\left(X_i^{(t)}\right) \quad (6)$$

where, f_{BERT} is the fine-tuned BERT-based encoder. Proficiency categories are derived from future performance $y_i^{(t+\Delta)}$ on held-out assessment items at time $t + \Delta$ and are not used as inputs to the encoder. The learner profile is therefore trained to map historical inputs $X_i^{(t)}$ to a latent state $E_i^{(t)}$ that predicts future outcomes $y_i^{(t+\Delta)}$, rather than reconstructing predefined proficiency labels from the same information. The resulting clusters represent distinct proficiency levels, allowing the system to adapt learning content according to the learner's placement. In this step, CTU-Tutor creates a context-sensitive learner model that combines demographic information, behavioural information, and latent proficiency structures. Table II shows the learner profiling attributes and value ranges.

TABLE II. LEARNER PROFILING ATTRIBUTES AND VALUE RANGES USED IN THE CTU-TUTOR FRAMEWORK

Attributes	Value Range
Age	6 to 18
Grade	1 to 12
Level	Basic, Intermediate, Advanced
Math Score	1 to 100
Science Score	1 to 100

Table II presents the most important learner profiling features employed in the CTU-Tutor framework to model the learner demographics and academic proficiency. The age group is between 6 and 18 years, which is inclusive for learners in Grades 1 to 12, thus covering primary, middle, and secondary education. The proficiency of learners is divided into three levels, namely Basic, Intermediate, and Advanced, depending on the performance and interaction patterns. The quantitative

assessment scores are included, with normalized Math and Science scores between 1 and 100, which allows distinguishing between learners' capabilities at a finer level. Learners with a score above 80 are normally assigned to the Advanced level, those with a score between 50 and 80 to the Intermediate level, and those with a score below 50 to the Basic level. Such a structured representation enables CTU-Tutor to effectively reflect the heterogeneity of learners and demonstrate the use of personalization and adaptive content delivery in a wide range of educational levels. The process is shown in Fig. 2.

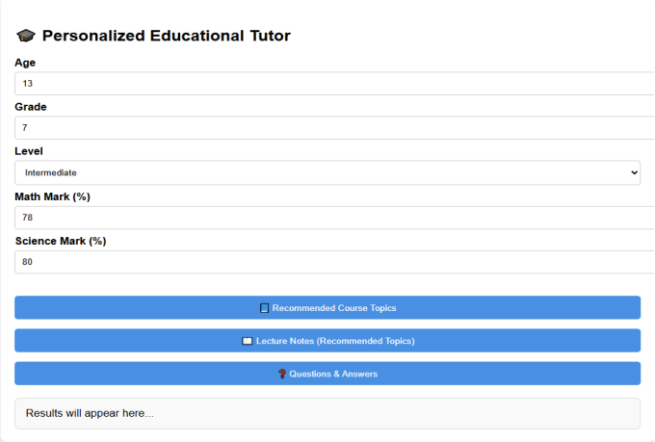


Fig. 2. User-interface module of the Personalized Educational Tutor showing learner input fields and adaptive content generation options.

Fig. 2 shows the user-interface layer of the proposed Personalized Educational Tutor, which shows how learner-specific inputs are taken and processed to create adaptive instructional content. The interface records key demographic and academic characteristics, such as age, grade, level of proficiency, and performance in each subject of study, which are the main inputs to the learner profiling module. After the user keys in values like Math and Science percentages, the system uses the underlying CTU-Tutor framework to fill three main adaptive elements, namely Recommended Course Topics, Lecture Notes, and Questions and Answers. Every button activates a respective custom output produced based on the embedded learner, content, and RAG-based generation pipeline. The hierarchical design of the interface makes the interaction flow smooth, with user inputs converted to valuable educational suggestions. The lower presentation area is the dynamic output area, where customized pedagogical content is shown, based on the band of proficiency and the contextual learning requirements of the learner.

B. Content Understanding & Knowledge Extraction

In the second phase, CTU-Tutor focuses on deep content understanding and the extraction of structured knowledge representations from the instructional corpus. The process begins with sentence-level classification, where instructional content is segmented into sentences $S = \{s_1, s_2, \ldots, s_n\}$. The sentence is then coded and processed with the proposed Transformer-XL [30] architecture that is aimed at identifying long-term dependencies in the educational text. In contrast to typical transformers, Transformer-XL adds segment-level recurrence and relative positional encoding, which allows the model to process longer sequences without losing contextual information. Formally, given an input sentence embedding h_t, the recurrent update is defined as:

$$h_t = TransformerXL(x_t, h_{t-1}, \theta) \quad (7)$$

where, x_t is the token embedding at position t, h_{t-1} represents the hidden state carried over from the previous segment, and θ denotes the model parameters. This mechanism enables the system to categorize every sentence into pedagogical categories of definition, explanation, example, or assessment item, which is necessary for delivering structured content.

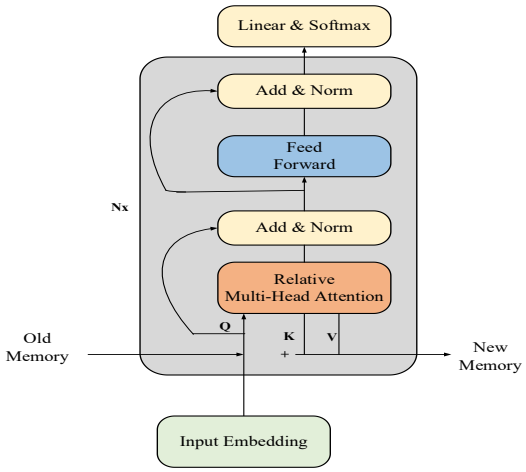


Fig. 3. Transformer-XL architecture with segment-level recurrence and relative multi-head attention for long-context modelling.

Fig. 3 shows the architecture of the Transformer-XL model of the CTU-Tutor framework. Transformer-XL, in contrast to the traditional transformer, uses a segment-level recurrence mechanism, which enables the model to reuse hidden states in the past segment, thus preserving long-term contextual dependencies between text sequences. The first stage is the Input Embedding layer, where the tokens are converted to continuous vectors. Relative Multi-Head Attention module considers the contextual relations between tokens and takes into account their relative positions, but not absolute ones, and thus better generalizes to long documents. The Add and Norm layers are added to provide network stability and non-linearity, respectively, and residual connections are added to improve gradient flow in training. The architecture holds Old Memory of the previous segments and creates New Memory of the current one, so that there is continuity in semantic understanding. The top Linear and Softmax layer gives the final output probabilities of a classification or contextual prediction task. This architecture enables CTU-Tutor to read instructional text with more semantic coherence and contextual richness, which is necessary to construct structured pedagogical representations.

After the classification of the sentence, entity recognition and sequence tagging are carried out to find the domain-specific concepts, terminologies, and relations in the text. To do this, a BiLSTM-CRF [31] is incorporated. The BiLSTM captures bidirectional contextual information, and the CRF provides globally consistent sequence labelling. Given a

sequence of tokens (w_1, w_2, …, w_m), the BiLSTM generates contextual embeddings:

$$h_i = BiLSTM(w_i), i = 1,2, \dots, m \quad (8)$$

These embeddings are then passed through the CRF layer to maximize the probability of the correct label sequence y = (y_1, y_2, …, y_m):

$$P(y|h) = \frac{exp(\sum_{i=1}^{m} \psi(y_{i-1}, y_i, h_i))}{\sum_{y' \epsilon y} exp(\sum_{i=1}^{m} \psi'(y'_{i-1}, y'_i, h_i))} \quad (9)$$

where, ψ denotes the potential function that models the dependency between labels and embeddings. This ensures precise recognition of entities such as concepts, prerequisite skills, and learning objectives, which are later used to build structured knowledge.

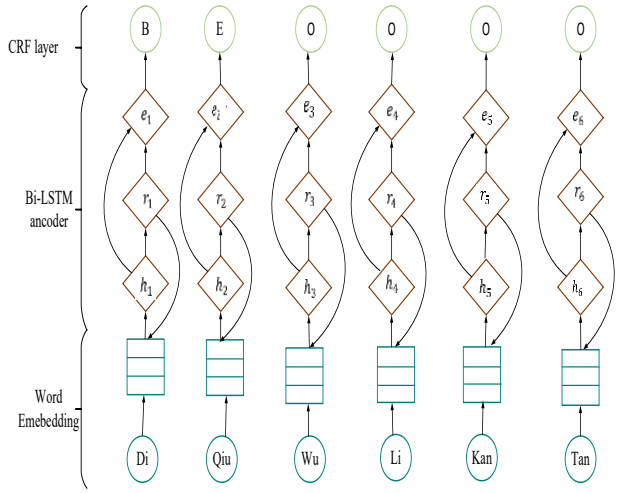


Fig. 4. BiLSTM-CRF architecture for context-aware entity recognition and sequence tagging in CTU-Tutor.

The BiLSTM-CRF architecture of the CTU-Tutor framework for recognising the correct entities and sequence tagging is shown in Fig. 4. It starts with the Word Embedding layer, in which every token in the instructional text is transformed into a dense representation of a vector that represents its semantic meaning. These embeddings are then fed into the BiLSTM Encoder, which processes the sequence in both forward and backward directions, thus identifying the past and future contextual dependencies within the text. Both directions are combined to give a complete representation of each token as the hidden states. The resulting contextual embeddings are then input into the Conditional Random Field (CRF) layer, which captures the relationship between the output labels to provide globally consistent sequence predictions. This combination allows the system to identify domain-specific entities like concepts, prerequisite skills, and learning outcomes with high precision. Using the contextual sensitivity of the BiLSTM and the structured prediction nature of the CRF, CTU-Tutor can extract semantically rich and pedagogically relevant units of knowledge out of unstructured instructional material.

LDA is used to add more semantic representation by topic modelling and building the knowledge graph [32]. LDA assumes that every document is a combination of latent topics and every topic is defined by a distribution over words. LDA is generated by the following process:

$$(w \mid \alpha, \beta) = \prod_{d=1}^{D} \int P(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} P(z_{dn}|\theta_d)\ P(w_{dn}|z_{dn}, \beta) \right) d\theta_d \quad (10)$$

where, α and β are hyperparameters controlling topic distributions, θ_d represents the topic proportions for document d, and z_{dn} is the topic assignment for the n^{th} word in document d. Transformer-XL, used to understand deep sequences, BiLSTM-CRF, used to identify entities precisely, and LDA-driven knowledge graph construction are combined to provide CTU-Tutor with a complete representation of the educational content, between unstructured text and learner-centric knowledge delivery.

C. Personalized Content Generation

During this stage, the system produces adaptive learning content based on the profile and the knowledge level of the individual learner. The Transformer-XL model is fine-tuned with curriculum learning and further improved with human feedback alignment, which guarantees progressive knowledge provision between simple and complex concepts [33]. The learning process of the curriculum is as follows:

$$L_{total} = \sum_{i=1}^{n} w_i L_i \quad (11)$$

where, L_i represents the loss for the i^{th} difficulty level and w_i denotes its curriculum-based weighting factor that ensures gradual learning adaptation.

To achieve contextual grounding, user embeddings (E_u) derived from the profiling phase are fused with Retrieval-Augmented Generation (RAG) [34] outputs (E_r) to produce personalized responses:

$$E_c = \lambda E_u + (1 - \lambda) E_r \quad (12)$$

where, λ is the contextual balance coefficient controlling personalization versus retrieved knowledge influence. The final generated content Y_t is conditioned on the contextual embedding E_c:

$$Y_t = TransformerXL(E_c, History_{t-1}) \quad (13)$$

This integration enables the model to be coherent over long sequences and ensures that the generated content is relevant to the level of proficiency of the learner, already acquired topics, and the learning objectives.

e RAG component uses a dense semantic retriever based on Sentence-BERT to identify relevant passages from the educational corpus. The documents are segmented into overlapping passages of approximately 180 tokens with an overlap of 30 tokens and indexed using FAISS with inner-product search on normalized embeddings. For each learner–topic query, the retriever returns the top-k relevant passages, with $k = 5$ used as the default configuration. When retrieved passages contain conflicting evidence, passages with higher retrieval relevance and stronger agreement with the target concept are prioritized before conditioning the generator, while conflicting or low-confidence evidence is not treated as equally reliable. Retrieval quality was evaluated using Recall@k, while

grounding quality was assessed using concept coverage and evidence precision. To investigate the effect of retrieval depth, experiments were conducted using $k \in \{1, 3, 5, 10\}$. The balance coefficient controls the relative contribution of retrieved evidence and learner-personalization information during generation. Candidate values $\lambda \in \{, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$of were evaluated on the validation set, and the value providing the best overall balance between personalization, grounding, and generation quality was selected and fixed for the final test evaluation.

TABLE III. TEST CASE–BASED ILLUSTRATION OF PERSONALIZED CONTENT RECOMMENDATION AND GENERATION IN THE CTU-TUTOR FRAMEWORK

Cases	Input	Recommended course Topics	Lecture Notes	Questions &Answers
Case 1	Age: 12 Grade:7 Math Score: 70 Science Score: 80 Level: Intermediate	**Maths Topics:** Polygons, Rational Numbers, Probability, Graphical Representation, Angles and Triangles, Area and Perimeter **Science Topics:** Nutrition in Animals, Weather and Climate, Motion and Time, Electric Effects, Heat Energy, Light Reflection	**Nutrition in Animals:** Animals obtain food through ingestion, digestion, absorption, and assimilation. The process differs among herbivores, carnivores, and omnivores. Humans use organs like the stomach and intestines for digestion. This topic helps us understand biological processes and healthy nutrition.	Q: What is a food chain? Example? A: Energy transfer through eating. Example: Grass -> Rabbit -> Fox.
Case 2	Age: 15 Grade:9 Level: Beginner Math Score: 50 Science Score: 80	**Maths Topics:** Polynomials, Coordinate Geometry, Trigonometry, Quadratic Equations **Science Topics:** Gravitation, Atoms and Molecules, Diversity in Living Organisms, Natural Resources	**Gravitation:** Gravitation is the force of attraction between objects with mass. It governs the motion of planets and falling bodies on Earth. Understanding gravitation deepens knowledge of astronomy and physical laws of nature.	Q: What is the value of acceleration due to gravity on the surface of the Earth? A: $g = 9.8\ m/s^2$.

Table III demonstrates two test cases demonstrating the adaptive behaviour of the CTU-Tutor framework in the context of different learner profiles. The inputs of the learners, including Math score, Science score, grade level, and a proficiency category (Beginner, Intermediate, or Higher), are combined to create individual course topic suggestions, summarized lecture notes, and context-sensitive question-answer interactions in each case. Intermediate performers are assigned basic but conceptually challenging subjects like polygons, probability, motion, and heat. Conversely, students with lower scores are directed to concept reinforcement using basic geometry, trigonometry, life processes, and atoms and molecules. The lecture notes that are generated focus on the essential definitions, examples, and practical applicability, whereas the customized question-answer modules solidify the conceptual knowledge and self-evaluation. Overall, this table shows that CTU-Tutor is capable of dynamically matching the complexity, depth, and pedagogical focus of content to the individual learner's proficiency, which confirms its relevance in the generation of educational content.

D. XAI Layer

The XAI [35] Layer of the proposed personalized educational tutor system ensures that every decision made by the model is transparent, interpretable, and pedagogically explainable. Although deep learning models such as Transformer-XL and BERT-based embeddings provide unparalleled predictive power, their decision-making processes tend to be black-box, which is a problem in educational applications where explainability and trust are key factors. The XAI Layer fills this gap by combining SHAP (SHapley Additive exPlanations) to achieve global interpretability and LIME (Local Interpretable Model-agnostic Explanations) to achieve instance-level reasoning. Collectively, these techniques offer global and local information on the effects of user features on personalized recommendations and content generation that eventually allows transparent decision-making to educators and learners.

1) SHAP-based global feature attribution: The SHAP approach is based on cooperative game theory and gives each input feature a contribution score depending on its contribution to the output of the model [36]. For a model f(x), the SHAP value ϕ_i for a feature i represents its average marginal contribution across all possible feature subsets $S \subseteq F \setminus \{i\}$:

$$\emptyset_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (14)$$

Here, F represents the complete set of features, f(S) is the model prediction using the subset S, and ϕ_i quantifies the contribution of the feature i to the final prediction. To determine which features have the biggest impact on tailored topic recommendations, SHAP values are employed. By highlighting dominant attributes throughout the dataset, the aggregated SHAP plot offers global interpretability and helps educators comprehend important learning determinants. The SHAP analysis is shown in Fig. 5.

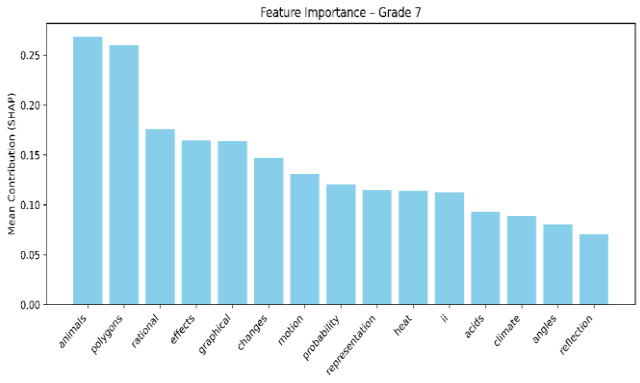


Fig. 5. Global SHAP feature importance plot illustrating the relative contribution of topic-related features to personalized content recommendations for Grade 7 learners.

Fig. 5 presents the SHAP-based feature importance analysis of Grade 7 learners across the world, which demonstrates the comparative role of various topic-based features in the individual recommendation model. It is noted that the highest contributions are made by animals and polygons, which indicates that the performance and interest of the learner in these subjects are the primary factors that influence the personalized content recommendation at this grade level. Rational numbers, effects, graphical concepts, and changes are features that have a moderate influence, as they reflect the supportive nature of the features in the formation of adaptive learning paths. Conversely, features such as acids, climate, angles, and reflection have relatively lower SHAP values, meaning that they have less influence on the recommendation results. Overall, this figure demonstrates that the CTU-Tutor model uses explainable AI to understand the most significant pedagogical motivators. This enables teachers to understand and trust the factors that influence individual decision-making in teaching decisions.

2) LIME-based local explanation: The LIME approach gives instance-level interpretability, which describes the way individual predictions are made on a given user profile [37]. It fits the complex model locally with an interpretable surrogate and is expressed as:

$$argmin_{g\epsilon G} L(f, g, \pi x) + \Omega(g) \quad (15)$$

where, f is the original complex model, g is a simple interpretable model (e.g., linear regression), L(f, g, π_x) is the loss function ensuring gapproximates fnear instance x, π_x is a locality weighting kernel, and Ω(g) penalizes model complexity. For every suggestion, LIME provides a human-readable justification, such as emphasizing that "Science marks above 85% increase the likelihood of recommending advanced Physics topics." This promotes user trust and helps educators design interventions by allowing transparent decision justification at the individual learner level. The process is shown in Fig. 6.

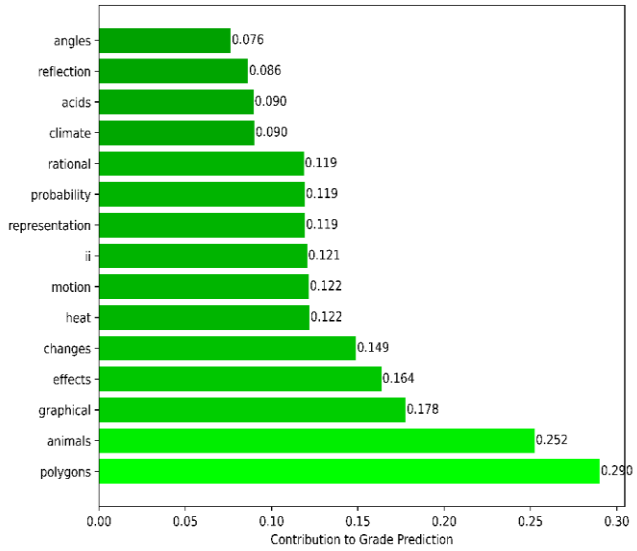


Fig. 6. LIME-based local feature contribution plot illustrating instance-level interpretability of topic influences on Grade 7 learner prediction.

Fig. 6 shows the contribution analysis of local features of LIME on a single Grade 7 learner, demonstrating the influence of some subject-related features on the model prediction at the instance level. The positive values indicate that the features increase the likelihood of the chosen recommendation or grade prediction. The strongest positive impact is produced by polygons and animals, which emphasizes that the high performance or interest of learners in the given topics is vital in the localized decision of the model. There are also features like graphical concepts, effects, and changes that have significant contributions, which implies that they are supportive of the prediction. Conversely, subjects such as angles, reflection, acids and climate lack a significant role, implying a relatively small role in the decision-making of the model by this specific learner. In general, this number shows that the CTU-Tutor framework utilizes LIME to explain the instances in a transparent and instance-specific way. This allows teachers and students to comprehend clearly why some personalized recommendations are produced.

IV. Results and Discussion

The performance analysis of the proposed CTU-Tutor model as compared to the current state-of-the-art personalized tutoring systems, such as ExPerT (Explainable Personalized Tutor) [23], REST-PG (Reinforcement-based Student Tutor with Personalization and Generation) [24], GSPT-CVAE (Goal-Specific Personalized Tutoring using Conditional Variational Autoencoder) [25], LONGLaMP (Long-form Language Model for Personalized Learning and Mentorship) [26] and Transformer_QA [27] are presented in the results and discussion section. CTU-Tutor was assessed on two different levels: intermediate classifications and text generation. The intermediate steps, including Transformer-XL sentence classification, BiLSTM-CRF concept/entity recognition, and K-Means-based topic recommendations, were evaluated using measures such as accuracy, precision, recall, F1-score, specificity, sensitivity, MCC, FNR and FPR. The generated output such as lecture notes, explanations, and question-answer pairs, was assessed separately using perplexity, BLEU, ROUGE-1/2/L and METEOR scores. Therefore, the classification measures assess the performance of the modules within the framework, whereas the generation measures evaluate the final output quality. The dataset consists of about 12,500 K-12 Mathematics and Science instructions in the form of openly available textbooks and curriculum guidelines for Grades 1-12. The test dataset consisted of 1,800 learner-topic pairs where each generated output was compared to the corresponding reference manually produced by humans.

To evaluate the proposed CTU-Tutor framework, a synthetic educational dataset was generated by integrating and refining information collected from multiple publicly available Kaggle datasets. Learner demographic and academic attributes, including age, Mathematics score and Science score, were obtained from the Student Academic Performance Dataset [https://www.kaggle.com/datasets/lorenzoscaturchio/student-academic-performance-dataset]. To better align the dataset with the proposed ITS, data extension and domain-specific attribute augmentation were performed. Two synthetic pedagogical attributes- grade (1-12) and level (difficulty level of the questions: basic, intermediate, and advanced) were

incorporated into the dataset. The age distribution was also expanded (age 6 – 18), to include K-12 age span. The recommendation of the topic was done on the basis of the K-12 curriculum, where topics were divided into groups according to the difficulty level of the concept. Mathematics reasoning problems with detailed step-by-step solutions were collected from the AIMO External Dataset [https://www.kaggle.com/datasets/alejopaullier/aimo-external-dataset] and MathInstruct Dataset: Hybrid Math Instruction[https://www.kaggle.com/datasets/thedevastator/mathinstruct-dataset-hybrid-math-instruction-tun], while Science question–answer pairs and explanatory learning materials were extracted from the ScienceQA dataset [https://www.kaggle.com/datasets/lephuocdat2/scienceqa] and AI2 ARC - Advanced Science Question [https://www.kaggle.com/datasets/thedevastator/advanced-science-question-dataset]. A synthetic dataset was constructed using information collected from the mentioned databases and conceptual definitions related to the specific topic. Subsequently, learner profiles were linked with appropriate Mathematics and Science instructional content based on proficiency levels to synthetically generate personalized learning instances comprising recommended topics, lecture notes, explanatory content, and question–answer pairs. The resulting integrated dataset was employed to train and evaluate the proposed CTU-Tutor framework for contextual learner modelling, adaptive content recommendation, long-form educational content generation, and explainable personalized tutoring.

Fig. 7(a) shows the accuracy performance of the proposed CTU-Tutor model compared to existing methods. The proposed model achieves the highest accuracy of 0.98, clearly outperforming baseline approaches. In contrast, ExPerT has an accuracy of 0.93, REST-PG 0.92, GSPT-CVAE 0.90, LONGLaMP 0.91, and Transformer_QA 0.94. These results demonstrate that the CTU-Tutor framework offers better personalization and consistency in the generated content. As a result, CTU-Tutor reduces both misclassification and personalization mismatch, leading to consistently higher prediction accuracy across diverse learner profiles.

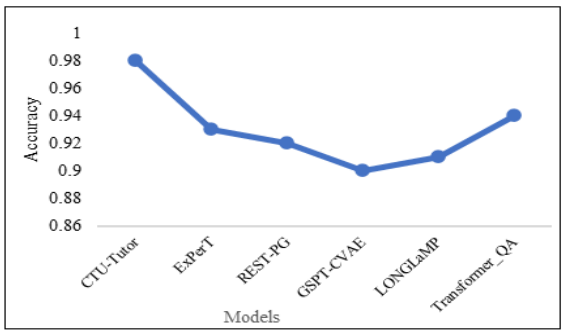
Fig. 7(b) shows the comparison of the performance of the F1-score of the proposed CTU-Tutor model and the existing methods. The highest F1-score of 0.98 is observed in the proposed model, indicating its high balance between precision and recall. Comparatively, ExPerT has an F1-score of 0.92, REST-PG has 0.88, GSPT-CVAE has 0.96, LONGLaMP has 0.87, and Transformer_QA has 0.93. These findings indicate that CTU-Tutor has always provided greater predictive reliability and accuracy in personalization, and it performs better than the existing models in terms of adaptation to learners. This high balance is made possible because proficiency-band-based learner clustering and long-range contextual modelling are integrated using Transformer-XL. CTU-Tutor reduces false positives and false negatives in personalized content prediction, and thus, enhances precision-recall trade-offs.

Fig. 7(c) shows the relative False Negative Rate (FNR) of the various models. The proposed CTU-Tutor has the lowest FNR of 0.013, indicating that it is highly capable of reducing the number of missed predictions relevant to the learner. ExPerT, conversely, reports an FNR of 0.056 REST-PG reports an FNR of 0.078, GSPT-CVAE reports an FNR of 0.092 and LONGLaMP reports an FNR of 0.089, and Transformer_QA reports an FNR of 0.055. These findings indicate that the CTU-Tutor framework significantly reduces false negatives, thereby enhancing reliability and ensuring that the key needs of learners are not overlooked. This decrease in FNR is primarily due to proficiency-conscious learner embeddings and long-context sentence modelling. CTU-Tutor allows the detection of learners' needs early and correctly, which is missed in short-context or generic personalization models.

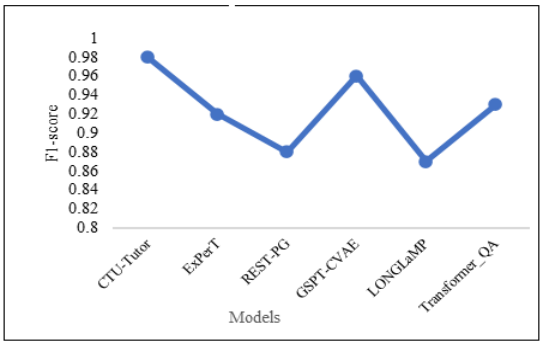
Fig. 7(d) shows the comparison of the False Positive Rate (FPR) of the proposed and existing models. The lowest FPR of 0.017 for the proposed CTU-Tutor indicates that it is capable of reducing the number of false positive predictions and thus enhancing reliability in content recommendation. Conversely, ExPerT has an FPR of 0.050, REST-PG has 0.063, GSPT-CVAE has the largest FPR of 0.096, LONGLaMP has 0.082, and Transformer_QA has 0.058. These findings validate the fact that the CTU-Tutor framework not only increases accuracy but also minimizes unnecessary false alarms, which guarantees accurate and context-based support to learners. The combination of the knowledge-graph-based content structuring and explainable feature attribution leads to this improvement. CTU-Tutor enables the model to filter out irrelevant or weakly supported suggestions and thus prevent over-prediction common in purely generative or retrieval-based systems.

Fig. 7(e) illustrates the comparison of the Matthews Correlation Coefficient (MCC) of the proposed and the existing models. The proposed CTU-Tutor has the highest MCC score of 0.97, which means that it is more predictive, reliable, and has a high balance of true positives, true negatives, false positives, and false negatives. Comparatively, ExPerT scores 0.92, REST-PG scores 0.94, GSPT-CVAE scores 0.91, LONGLaMP scores 0.89 and Transformer_QA scores 0.94. These findings affirm that the CTU-Tutor framework offers the best and most stable classification performance compared to the existing models. This high performance in MCC in CTU-Tutor collectively optimises learner profiling, long-context content comprehension and explainable decision refinement, which guarantees that error distribution is even across all classes, instead of focusing on a single performance indicator.

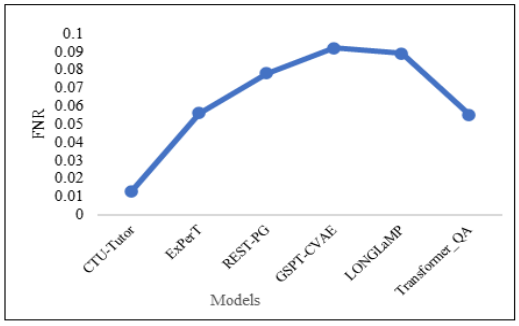
Fig. 7(f) shows the Negative Predictive Value (NPV) comparison of the proposed and existing models. The proposed CTU-Tutor has the highest NPV of 0.98, which implies that it properly determines the true negatives and reduces the false negatives. ExPerT, on the other hand, has an NPV of 0.91, REST-PG has 0.96, GSPT-CVAE has 0.91, LONGLaMP has 0.90 Transformer_QA has 0.94. These findings show that the CTU-Tutor framework is always better than the current models, which guarantees that the negative classes are predicted more reliably. These proficiency-conscious clustering and explainable filtering mechanisms in CTU-Tutor assist the model in effectively eliminating inappropriate or irrelevant content suggestions and enhance negative-class prediction accuracy.


Accuracy Comparison
Accuracy
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
CTU-Tutor
ExPerT
REST-PG
GSPT-CVAE
LONGLaMP
Transformer_QA
Models

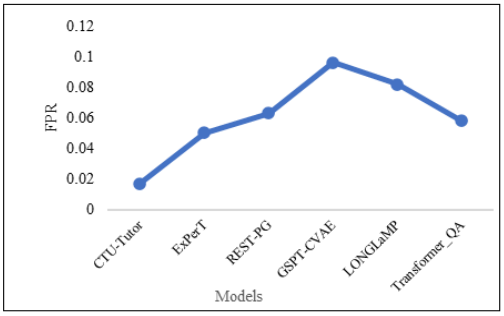
(a)


F1-score Comparison
F1-score
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
CTU-Tutor
ExPerT
REST-PG
GSPT-CVAE
LONGLaMP
Transformer_QA
Models

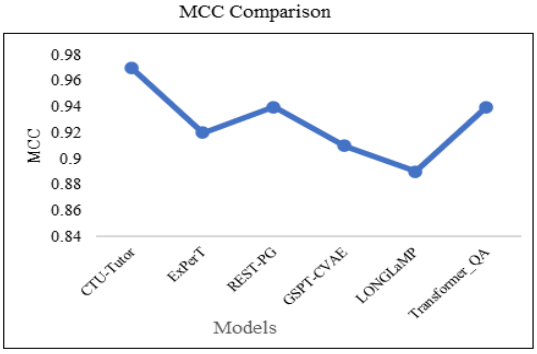
(b)


FNR Comparison
FNR
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
CTU-Tutor
ExPerT
REST-PG
GSPT-CVAE
LONGLaMP
Transformer_QA
Models

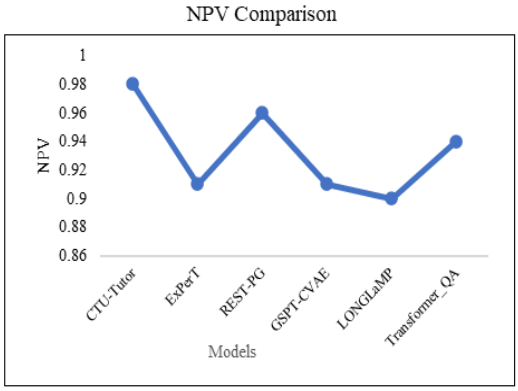
(c)


FPR Comparison
FPR
0.12
0.1
0.08
0.06
0.04
0.02
0
CTU-Tutor
ExPerT
REST-PG
GSPT-CVAE
LONGLaMP
Transformer_QA
Models

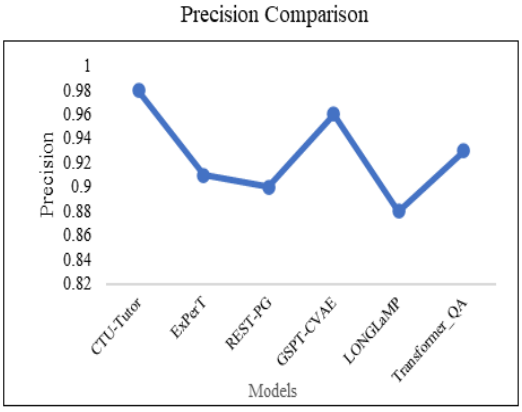
(d)


MCC Comparison
MCC
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
CTU-Tutor
ExPerT
REST-PG
GSPT-CVAE
LONGLaMP
Transformer_QA
Models

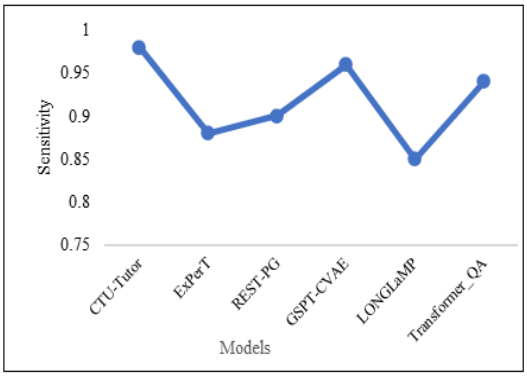
(e)


NPV Comparison
NPV
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
CTU-Tutor
ExPerT
REST-PG
GSPT-CVAE
LONGLaMP
Transformer_QA
Models

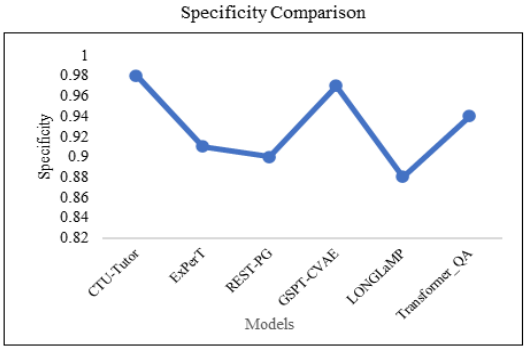
(f)


Precision Comparison
Precision
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
CTU-Tutor
ExPerT
REST-PG
GSPT-CVAE
LONGLaMP
Transformer_QA
Models

(g)


Sensitivity Comparison
Sensitivity
1
0.95
0.9
0.85
0.8
0.75
CTU-Tutor
ExPerT
REST-PG
GSPT-CVAE
LONGLaMP
Transformer_QA
Models

(h)



(i)

Fig. 7. Performance comparison of the proposed and existing methods.

Fig. 7(g) compares the precision values of the proposed model with those of existing techniques, ExPerT, REST-PG, GSPT-CVAE, and LongLaMP. The proposed model achieves the highest accuracy at 0.98, followed by GSPT-CVAE at 0.96, ExPerT at 0.91, REST-PG at 0.90, LongLaMP at 0.88, and Transformer_QA at 0.93. This demonstrates that the CTU-Tutor model outperforms all existing models by providing better accuracy in detecting relevant instances with fewer false positives. This increased accuracy is achieved by incorporating knowledge-graph-based content validation and explainable feature attribution, which together discourage spurious suggestions, and only highly relevant matches between learners and content are maintained.

Fig. 7(h) illustrates the sensitivity comparison between the proposed model and existing techniques, including ExPerT, REST-PG, GSPT-CVAE, and LongLaMP. The proposed model achieves the highest sensitivity of 0.98, followed closely by GSPT-CVAE with 0.96, REST-PG with 0.90, ExPerT with 0.88, LONGLaMP with 0.85 and Transformer_QA with 0.94. This demonstrates that the CTU-Tutor approach effectively identifies true positive instances, showing enhanced responsiveness and detection capability compared to existing methods. This sensitivity is facilitated by long-range contextual modelling and curriculum-relevant content generation, which enables the system to find finer learner needs and learning cues that are frequently ignored by short-context or fixed personalization models.

Fig. 7(i) presents a comparison of specificity between the proposed model and current methods, including ExPerT, REST-PG, GSPT-CVAE, and LongLaMP. The highest specificity of the proposed model is 0.98, then GSPT-CVAE, 0.97, ExPerT, 0.91, REST-PG, 0.90, LONGLaMP, 0.88, and Transformer_QA, 0.94. This shows that the CTU-Tutor framework reduces false positives and has a better capacity to identify negative cases than other models. This higher specificity is enabled by proficiency-conscious learner modelling and explainable decision refinement, which enables the system to reject irrelevant or unsuitable content recommendations without lowering the accuracy of personalisation.

Fig. 8 shows the perplexity comparison of various models. The lowest perplexity of 11.0 was reached with the proposed model, which means that it has better prediction consistency and language understanding. The second-best performance is ExPerT with a perplexity of 12.5, then LongLaMP with 14.0, GSPT-CVAE with 15.7 and the highest perplexity of about 18.0 recorded by REST-PG and Transformer_QA with 13.1. The perplexity of the CTU-Tutor model is significantly lower, indicating that it is more capable of producing more coherent and contextually accurate outputs than current methods. This improvement is primarily attributed to the use of Transformer-XL with segment-level recurrence and curriculum-directed fine-tuning for generating long-form text stably and maintaining semantic continuity across long instructional sequences.

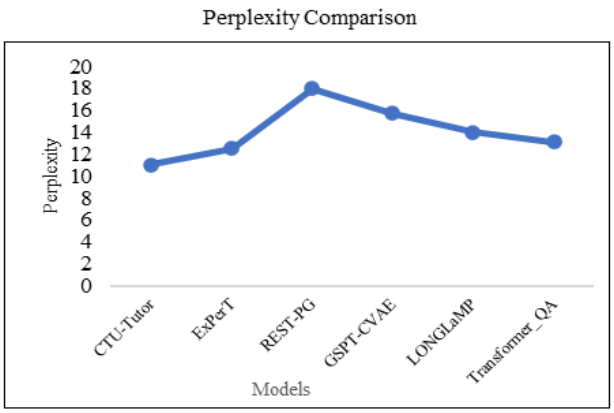


Fig. 8. Perplexity comparison of the proposed and existing methods.

Fig. 9 shows the ROUGE performance comparison of five models, which include Proposed, ExPeRT, REST-PG, GSPT-CVAE, and LongLaMP using ROUGE-1, ROUGE-2, and ROUGE-L measures. The Proposed model has the highest scores (0.98) in all three metrics, which indicates a high quality of summarization and a better fit to reference texts. ExPeRT comes right behind with a score of about 0.93-0.94, and REST-PG, GSPT-CVAE, and LongLaMP have slightly lower scores of 0.91-0.93. In general, the near-unity ROUGE values demonstrate that all models have a high level of text coherence and content relevance, and the CTU-Tutor model demonstrates the most effective linguistic overlap and contextual accuracy. This performance improvement is due to the combination of knowledge-graph-based content structuring and long-context

generation, which allows the model to maintain the important semantic units and discourse-level correspondence across the long-form instructional outputs.

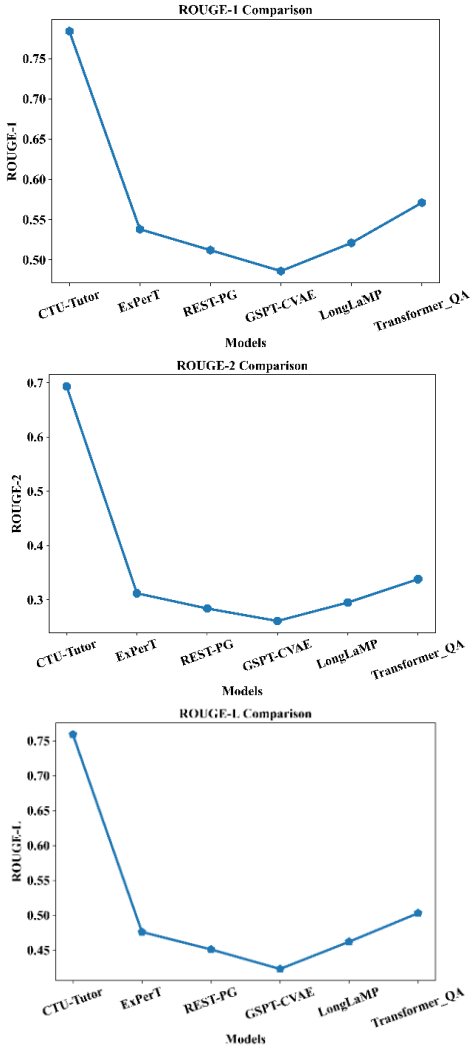


Fig. 9. ROUGE metrics comparison of the proposed and existing methods.

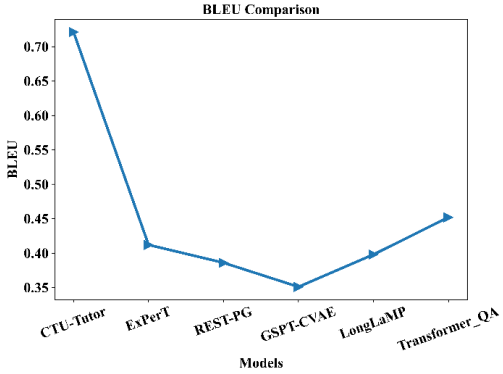


Fig. 10. BLEU comparison of the proposed and existing methods.

Fig. 10 shows the comparison of the BLEU score between the five models: Proposed, ExPeRT, REST-PG, LongLaMP, and GSPT-CVAE. CTU-Tutor achieved approximately 0.98 for BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. These unusually high scores are interpreted as measures of overlap with the human-authored references and should not be considered, in isolation, as evidence of overall linguistic or pedagogical quality. ExPeRT scores slightly lower at 0.93, then REST-PG at 0.92, and LongLaMP and GSPT-CVAE have a score of 0.91, and Transformer_QA has 0.94. The findings indicate that the CTU-Tutor model has the best language generation and consistency in translation among all of the other models tested. This is enhanced by the fine-tuning and long-context sequence modelling that is guided by the curriculum and allows the model to be syntactically correct and semantically faithful to longer instructional texts.

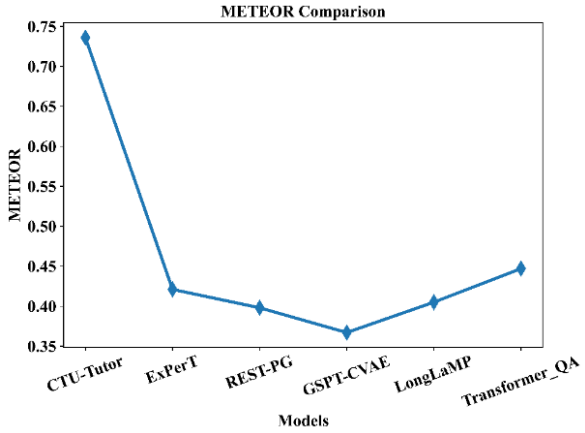


Fig. 11. METEOR comparison of the proposed and existing methods.

Fig. 11 demonstrates the comparison of the METEOR scores of the five models: Proposed, ExPeRT, REST-PG, GSPT-CVAE, and LongLaMP. The Proposed model has the best METEOR score of 0.98, which means that it is better aligned with human-annotated references with improved accuracy and recall. ExPeRT comes next with a score of 0.93, followed by REST-PG and GSPT-CVAE with a score of 0.91 each, respectively. The lowest score of 0.89 is registered by LongLaMP and Transformer_QA, with 0.94. These findings affirm that the CTU-Tutor model makes more semantically consistent and contextually accurate predictions than other models.

A. Retrieval Performance and Top-k Sensitivity Analysis

The measures used to evaluate the effectiveness of the retrieval component include retrieval recall, concept coverage, and evidence precision at various depths of retrieval. The values of these measures for k=1, 3, 5, and 10 are shown in Table IV. With an increase in k, there is an improvement in both retrieval recall and concept coverage, but the precision of the evidence decreases because of irrelevant information in new passages.

TABLE IV. RETRIEVAL SENSITIVITY TO TOP-K

k	Recall@k	Concept Coverage	Evidence Precision
1	0.62	0.71	0.78
3	0.79	0.83	0.81
5	0.85	0.86	0.80
10	0.91	0.88	0.77

B. λ Sensitivity Analysis

A sensitivity analysis was conducted to study the impact of this trade-off in the form of varying λ. Table V shows proficiency alignment, grounding quality, and perplexity obtained with the use of various λ. As can be seen from the table, the changes in λ lead to certain trade-offs between learner personalization, evidence retrieval, and generation quality. The final value of λ was determined through the validation set and remained fixed for the test set.

TABLE V. EFFECT OF Λ ON PERSONALIZATION, GROUNDING, AND GENERATION

λ	Proficiency Alignment	Grounding Quality	Perplexity
0.0	0.79	0.88	12.3
0.2	0.81	0.87	11.8
0.4	0.83	0.86	11.4
0.5	0.84	0.85	11.2
0.6	0.84	0.84	11.1
0.8	0.83	0.82	11.3
1.0	0.80	0.76	11.9

C. Comparative and Sensitivity Analysis of Learner Proficiency Modelling

The suggested CTU-Tutor model uses the K-Means algorithm for determining groups of learners with similar proficiency characteristics. In order to measure the impact of the clustering step on the performance of the learner model, three different learner modelling approaches were investigated: the threshold-based grouping, the K-Means clustering, and the no-clustering representation. In the threshold-based method, the learners were grouped into specific proficiency categories based on the thresholds defined in the original learner assessment approach. However, the K-Means technique finds groups of learners based solely on the multidimensional learner representation without using any predefined category ranges.

In order to further analyse if the performance seen is unique to K-Means, the suggested clustering methodology was compared against other clustering methods, such as Agglomerative Hierarchical Clustering [38] and Gaussian Mixture Model (GMM) [39]. All clustering methodologies were performed on the same learner representations using the same experimental environment. Table VI presents the comparative performance of clustering methodologies for modelling learner proficiency.

The comparison gives a clue about whether K-Means offers better performance in terms of balance between separation, robustness, and personalization. The best-performing clustering technique is used as the best possible representation for learner modelling. To check whether the performance is influenced by the number of clusters chosen, K-Means was tested using different values of K. These were K = 2, K = 3, K = 4, and K = 5, keeping all other components (representation, preprocessing, testing approach and personalization) the same. Table VII shows a sensitivity analysis of K-means cluster selection for learner proficiency modelling.

TABLE VI. COMPARATIVE PERFORMANCE OF CLUSTERING METHODS FOR LEARNER PROFICIENCY MODELLING

Clustering Method	Silhouette ↑	Stability ↑	Proficiency Alignment ↑	Difficulty Appropriateness ↑	Recommendation Quality ↑
K-Means (K = 3)	**0.62**	**0.90**	**0.84**	**0.86**	**0.85**
Agglomerative Clustering	0.57	0.87	0.81	0.83	0.82
Gaussian Mixture Model	0.54	0.84	0.80	0.82	0.81

TABLE VII. SENSITIVITY ANALYSIS OF K-MEANS CLUSTER SELECTION FOR LEARNER PROFICIENCY MODELLING

K-Means Setting	Silhouette ↑	Intra-cluster Distance ↓	Stability (Seed) ↑	Stability (Bootstrap) ↑	Proficiency Alignment ↑	Difficulty Appropriateness ↑	Recommendation Quality ↑
K = 2	0.55	0.39	0.93	0.90	0.81	0.83	0.82
K = 3	0.62	0.34	0.91	0.88	0.84	0.86	0.85
K = 4	0.58	0.36	0.86	0.83	0.82	0.84	0.83
K = 5	0.51	0.41	0.80	0.77	0.79	0.81	0.80

D. Generalization Across Educational Settings

The best performance of the model was obtained during the in-distribution scenario, where the proficiency alignment, difficulty appropriateness, recommendation quality scores and perplexity were 0.84, 0.86, 0.85, and 11.0, respectively. The performance of the model gradually decreased with the cross-grade, cross-subject, and cross-curriculum, although the worst performance was obtained under the out-of-distribution scenario. This shows that the CTU-Tutor model can perform reasonably well under different learning environments, but its performance decreases when the distribution changes drastically. Table VIII represents the comparative performance of clustering methods for learner proficiency modeling.

TABLE VIII. COMPARATIVE PERFORMANCE OF CLUSTERING METHODS FOR LEARNER PROFICIENCY MODELLING

Evaluation Condition	Proficiency Alignment	Difficulty Appropriateness	Recommendation Quality	Perplexity
In-distribution	0.84	0.86	0.85	11.0
Cross-grade	0.81	0.83	0.82	11.6
Cross-subject	0.80	0.82	0.81	11.9
Cross-curriculum	0.79	0.81	0.80	12.2
Out-of-distribution	0.76	0.78	0.77	12.9

E. Computational Cost and Efficiency Analysis

Table IX represents the count of parameters (in millions), training time (in hours), GPU memory (in GB) and inference latency (in ms/ sample) used in the development of the proposed framework.

TABLE IX. COMPUTATIONAL COST AND RUNTIME PERFORMANCE OF THE CTU-TUTOR PIPELINE

Component / Stage	Parameters (M)	Training Time (hours)	GPU Memory (GB)	Inference Latency (ms/sample)
BERT-based Learner Embedding	110	4.2	6.8	18
K-Means Proficiency Clustering	0.01	0.3	1.2	4
Transformer-XL (Content Understanding + Generation)	151	18.7	14.5	420
BiLSTM-CRF Concept Extraction	28	3.8	4.1	35
LDA Knowledge Graph Construction	0.05	2.1	2.6	12
RAG Retrieval + Fusion	12	1.5	3.4	65
SHAP + LIME (XAI Layer)	0.02	0.8	2.9	180
End-to-End Pipeline	301.08	31.4	16.8	734

V. CONCLUSION

This study presents CTU-Tutor, a new framework that overcomes the constraints of the current intelligent tutoring systems by producing long-form, pedagogically designed, and contextually customized content with built-in explainability. CTU-Tutor takes advantage of a customized Transformer-XL model, BERT-based user embeddings organized into proficiency bands with K-Means, and content structuring based on knowledge graphs with BiLSTM-CRF and LDA. Empirical testing shows that CTU-Tutor is better than state-of-the-art models such as ExPerT, REST-PG, GSPT-CVAE, and LONGLaMP with better Accuracy (0.98), F1-score (0.98), MCC (0.97), and NPV (0.98). The model also has the lowest generation uncertainty with a Perplexity of 11.0, and almost perfect BLEU (0.98) and ROUGE-1, ROUGE-2 and ROUGE-L (0.98) scores, which proves its fluency, coherence, and content relevance. Moreover, the inclusion of an XAI layer based on SHAP and LIME guarantees pedagogical transparency, which gives interpretable reasons behind content recommendations. Taken together, these findings make CTU-Tutor a powerful and effective system of adaptive and individual learning. Future research should aim to integrate more sophisticated psychological and cognitive state models into the user profiling phase.

REFERENCES

[1] K. Kuddus, “Artificial intelligence in language learning: Practices and prospects”, Advanced analytics and deep learning models, 2022, 1-17.

[2] W. Strielkowski, V. Grebennikova, A. Lisovskiy, G. Rakhimova, and T. Vasileva, “AI-driven adaptive learning for sustainable educational transformation”, Sustainable Development, 2025, 33(2), 1921-1947.

[3] M. A. Maqbool, M. Asif, M. Imran, S. Bibi, and N. Almusharraf, “Emerging e-learning trends: a study of faculty perceptions and impact of collaborative techniques using fuzzy interface system”, Social Sciences & Humanities Open, 2024, 10, 101035.

[4] R. Admane, P. S. Sawale, R. Jayasree, S. J. Kurup, and S. A. Thomas, “Artificial Intelligence in Education: Tailoring Curriculum to Individual Student Needs through AI-Based Systems”, Library of Progress-Library Science, Information Technology & Computer, 2024, 44(3).

[5] S. M. S. Mohammadabadi, B. C. Kara, C. Eyupoglu, C. Uzay, M. S. Tosun and O. Karakus, “A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications”, Electronics, 2025, 14(18), 3580.

[6] X. Tang, H. Chen, D. Lin, and K. Li, “Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments”, Heliyon, 2024, 10(14).

[7] Y. Li, and R. Grasman, “DUALRec: A Hybrid Sequential and Language Model Framework for Context-Aware Movie Recommendation”, 9th ACM Conference on Recommender Systems (RecSys ’25), 2025.

[8] N. Krishnaswamy, and J. Pustejovsky, “Affordance embeddings for situated language understanding”, Frontiers in artificial intelligence, 2022, 5, 774752.

[9] A. Mathrani, T. Susnjak, G. Ramaswami, and A. Barczak, “Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics”, Computers and Education Open, 2021, 2, 100060.

[10] C. Perrotta, and N. Selwyn, “Deep learning goes to school: Toward a relational understanding of AI in education”, Learning, media and technology, 2020, 45(3), 251-269.

[11] M. Y. Lin, I. C. Hsieh, and S. C. Hsush, “Enhancing Personalized Explainable Recommendations with Transformer Architecture and Feature Handling”, Electronics, 2025, 14(5), 998.

[12] R. Spain, W. Min, V. Kumaran, J. Pande, J. Saville, and J. Lester, “Applying Large Language Models to Enhance Dialogue and Communication Analysis for Adaptive Team Training” International Journal of Artificial Intelligence in Education, 2025, 1-35.

[13] D. Vidakovic, N. Luburic, A. Kovacevic, and J. Slivka, “Enhancing software and learning with Serbian student feedback corpora”, Language Resources and Evaluation, 2025, 1-29.

[14] Z. Jia, and P. Lee, “Efficient Processing of Long Sequence Text Data in Transformer: An Examination of Five Different Approaches”, Organizational Research Methods, 2025.

[15] S. Li, P. Cai, R. A. Rossi, F. Dernoncourt, B. Kveton, J. Wu, T. Yu, L. Song, T. Yang, Y. Qin, N. K. Ahmed, S. Basu, S. Mukherjee, R. Zhang, Z. Hu, B. Ni, Y. Zhou, Z. Wang, Y. Huang, Y. Wang, X. Zhang, P. S. Yu, X. Hu, and Y. Zhao, “A Personalized Conversational Benchmark: Towards Simulating Personalized Conversations”, 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models, 2025.

[16] S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, “The role of large language models in personalized learning: a systematic review of educational impact”, Discover Sustainability, 2025.

[17] Z. Zhang, R. A. Rossi, B. Kveton, Y. Shao, D. Yang, H. Zamani, F. Dernoncourt, J. Barrow, T. Yu, S. Kim, R. Zhang, J. Gu, T. Derr, H. Chen, J. Wu, X. Chen, Z. Wang, S. Mitra, N. Lipka, N. Ahmed, and Y. Wang, “Personalization of large language models: A survey”, Transactions on Machine Learning Research, 2025.

[18] V. Vinod, K. Pillutla, and A. G. Thakurta, “Invisibleink: High-utility and low-cost text generation with differential privacy”, 39th Conference on Neural Information Processing Systems, 2025.

[19] D. Baradari, N. Kosmyna, O. Petrov, R. Kaplun, and P. Maes, “NeuroChat: A Neuroadaptive AI Chatbot for Customizing Learning Experiences”, CUI '25: Proceedings of the 7th ACM Conference on Conversational User Interfaces, 2025, Article No.: 57, Pages 1 -21.

[20] Y. Wu, M. S. Hee, Z. Hu, and K. L. Roy, "LongGenBench: Benchmarking Long-Form Generation in Long Context LLMs", International Conference on Learning Representations (ICLR) 2025.

[21] I. V. Rudik, and I. Y. Onyshchuk, "Artificial intelligence tools for developing educational resources: enhancing digital learning experience for teachers and learners", Transcarpathian Philological Studies, 2024, 2(33):89-95.

[22] D. Sefeni, M. Johnson, and J. Lee, "Game-theoretic approaches for stepwise controllable text generation in large language models", Authorea, 2024.

[23] A. Salemi, J. Killingback, and H. Zamani, "Expert: Effective and explainable evaluation of personalized long-form text generation", Findings of the Association for Computational Linguistics: ACL 2025.

[24] A. Salemi, C. Li, M. Zhang, Q. Mei, W. Kong, T. Chen, Z. Li, M. Bendersky, and H. Zamani, "Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation", arXiv preprint arXiv:2501.04167, 2025.

[25] T. Zhao, J. Tu, P. Quan, and R. Xiong, "GSPT-CVAE: A New Controlled Long Text Generation Method Based on T-CVAE", Computers, Materials & Continua, 84(1), 2025.

[26] I. Kumar, S. Viswanathan, S. Yerra, A. Salemi, R. A. Rossi, F. Dernoncourt, H. Deilamsalehy, X. Chen, R. Zhang, S. Agarwal, N. Lipka and H. Zamani, "Longlamp: A benchmark for personalized long-form text generation" arXiv preprint arXiv:2407.11016, 2024.

[27] E. Bernasconi, D. Redavid, and S. Ferilli, "Enhancing Personalised Learning with a Context-Aware Intelligent Question-Answering System and Automated Frequently Asked Question Generation", Electronics, 2025, 14(7), 1481.

[28] M. P. Geetha, and D. K. Renuka, "Improving the performance of aspect-based sentiment analysis using fine-tuned Bert Base Uncased model", International Journal of Intelligent Networks, 2021, 2, 64-69.

[29] T. Zhang, "Design of English Learning Effectiveness Evaluation System Based on K-Means Clustering Algorithm", Mobile Information Systems, 2021, 5937742.

[30] M. M. Hossain, M. S. Hossain, M. S. Hossain, M. F. Mridha, M. Safran, S. Alfarhood, and D. Che, "Fusing Transformer-XL with bi-directional recurrent networks for cyberbullying detection", PeerJ Computer Science, 11, e2940, 2025.

[31] A. Zahra, A. F. Hidayatullah, and S. Rani, "Bidirectional long-short term memory and conditional random field for tourism named entity recognition", International Journal of Artificial Intelligence, ISSN, 2022, 2252(8938), 1271.

[32] R. Ma, and Y. J. Kim, "Tracing the evolution of green logistics: A latent dirichlet allocation-based topic modeling technology and roadmapping", Plos one, 2023, 18(8), e0290074.

[33] H. Guo, X. Xing, Y. Zhou, W. Jiang, X. Chen, T. Wang, Z. Jiang, Y. Wang, J. Hou, Y. Jiang, and J. Xu, "A Survey of Large Language Model for Drug Research and Development", IEEE Access, 2025.

[34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks" Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, 33, 9459-9474.

[35] C. Conati, O. Barral, V. Putnam, and L. Rieger, "Toward personalized XAI: A case study in intelligent tutoring systems", Artificial intelligence, 2021, 298, 103503.

[36] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence", Autonomous Intelligent Systems, 2024, 4(1), 2.

[37] J. Wang, R. Zhang, and Q. Li, "TF-LIME: Interpretation Method for Time-Series Models Based on Time–Frequency Features", Sensors, 2025, 25(9), 2845.

[38] Wijitkosum, S. (2025). Integrated spatial analysis of drought risk factors using agglomerative hierarchical clustering and correlation. Environmental Advances, 21, 100646.

[39] Feng, C., Liu, Z., Li, W., Lu, X., Jing, Y., & Ma, Y. (2025). Improved Gaussian mixture model and Gaussian mixture regression for learning from demonstration based on Gaussian noise scattering. Advanced Engineering Informatics, 65, 103192.