PBAT: A Profile-based Adaptive Testing Framework Powered by LLMs

Pankaj P. Mishra [1], V. Venkataramanan [1], Wen Gu [2,*], Keyur Patel [1], Tirth Patel [1], Asmi Moghe [1], and Adwait Patankar [1]

[1] Department of Information Technology, K J Somaiya School of Engineering (formerly K J Somaiya College of Engineering), Somaiya Vidyavihar University, Vidyavihar, Mumbai, 400077, Maharashtra, India
[2] Department of Computer Science and Engineering, Nagoya Institute of Technology, Gokiso, Showa, Nagoya, 466-8555, Aichi, Japan
Email: pankaj.mishra@somaiya.edu (P.P.M.); venkataramanan@somaiya.edu (V.V.); wgu@nitech.ac.jp (W.G.); keyurp@somaiya.edu (K.P.); airthp@somaiya.edu (T.P.); asmip@somaiya.edu (A.M.); adwaitp@somaiya.edu (A.P.)
*Corresponding author

Manuscript received October 15, 2025; revised December 1, 2025; accepted January 27, 2026; published July 15, 2026

***Abstract*—Artificial intelligence–driven adaptive learning systems are redefining modern education by enabling automated, personalized, and data-driven assessments. Existing quiz generation models still face two key challenges: hallucination, where Large Language Models (LLMs) produce factually incorrect questions, and limited personalization, which reduces the learning impact. To overcome these issues, this study introduces the Profile-based Adaptive Testing (PBAT) framework, which unifies Retrieval-Augmented Generation (RAG) for factual grounding, Item Response Theory (IRT) for psychometric calibration, and a Multi-Armed Restless Bandit (MARB) algorithm for adaptive sequencing. PBAT uses reinforcement-driven optimization to dynamically generate, validate, and adjust quizzes guided by learner profiles in real time. Comprehensive experiments on benchmark educational datasets demonstrate that PBAT achieves superior performance, reducing hallucination rates to 8.2%, increasing BLEU and F1-Scores, and improving learner retention by 15.6% after seven days compared with conventional IRT and static LLM-based systems. The framework also lowers cognitive load and maintains smooth difficulty transitions, improving engagement and fairness. Overall, the PBAT bridges the psychometric rigor of adaptive testing with the generative intelligence of LLMs, offering a scalable and pedagogically coherent solution for next-generation intelligent learning and assessment systems.**

***Keywords*—adaptive learning, question generation, personalized assessment**

I. Introduction

Assessment is one of the most important elements in both education and professional training. This structured process serves to assess learners' level of knowledge and enables instructional interventions to track their progress. Traditionally, learners have been assessed through static examinations, using the same sets of questions regardless of their previous experience or current level of expertise, dedication to the material, etc. As a result, novice learners may suffer from an overload of questions that they are unable to answer because of their unfamiliarity with the material, while experienced learners may be provided with questions that fail to push them to further develop their skills, thus causing them to feel less motivated and less able to learn effectively. The advent of Computerized Adaptive Testing (CAT), in addition to the proliferation of the use of Large Language Models (LLMs), has allowed for greater customization in assessments through the development of a dynamic sequencing of questions based on a learner's response, as well as an adaptive process for evaluating learners. CAT has strong psychometric backing, as CAT-based assessments typically use Item Response Theory (IRT) as the basis for their assessments; however, CAT frameworks depend on pre-calibrated question banks that have been developed through intensive labor and the expenditure of considerable resources [1–3]. Conversely, systems that develop questions using large language models create linguistically reliable and contextually substantial questions using deep neural networks [4–7]; however, these systems are still susceptible to the two main pitfalls associated with CAT-based frameworks: (i) generating hallucinations (i.e., questions generated that are conceptually or contextually inaccurate) [8] and (ii) pedagogical misalignment (i.e., failure to provide a question that is aligned with the learner's current level of development). Such generated items fail to follow cognitive progression frameworks such as Bloom's taxonomy [9] or learner performance trajectories. Moreover, most adaptive quiz systems still rely on heuristic selection mechanisms and lack integration with probabilistic decision-making approaches such as Reinforcement Learning (RL) or multi-Armed Banks (MABs) [10, 11]. These models can effectively balance the exploration of unmastered concepts and exploitation of known strengths, leading to better adaptivity and fairness. Despite their potential, few existing frameworks incorporate these methods into LLM-based question generation in a cohesive manner. To address these limitations, this paper introduces Profile-based Adaptive Testing (PBAT), a unified AI-driven framework that integrates Retrieval-Augmented Generation (RAG) for factual grounding, IRT for psychometric calibration, and a Multi-Armed Restless Bandit (MARB) model for adaptive sequencing. The RAG component grounds question generation in retrieved source content, improving contextual relevance and reducing hallucination risk [12, 13], whereas the IRT model provides statistical rigor in estimating learner ability. The MARB module functions as an adaptive decision engine that continuously selects the next question based on real-time learner performance. Its "restless" property enables each knowledge domain, represented as an arm, to evolve dynamically, even when not selected for questioning. By enabling the model to incorporate the natural processes of time and forgetting into their estimates of learner proficiency [14, 15], it will be possible to provide timely and accurate assessments of a learner's proficiency level. By balancing the time spent on developing weaker aspects of conceptual understanding with opportunities to reinforce through the use of already developed strengths, MARB is

doi: 10.18178/ijiet.2026.16.7.2641

able to facilitate gradual difficulty changes and higher levels of learner engagement and long-term adaptability. Collectively, these elements create a fact-based, pedagogically sound, and learner-focused adaptive assessment ecosystem.

Although prior work on adaptive assessment and AI-supported learning has advanced significantly, several critical challenges remain. First, traditional IRT-based CAT systems rely on large pre-validated static item banks, which are time consuming and expensive to develop, limiting scalability and cross-domain generalization. Second, although LLM-based quiz generators produce linguistically fluent questions, they often generate factually incorrect or irrelevant content because of the absence of retrieval grounding and psychometric validation. Third, most existing adaptive systems depend on heuristic rule mechanisms or fixed difficulty progression, and do not incorporate probabilistic or reinforcement-based approaches for adaptivity. Finally, prior studies predominantly evaluated such systems using linguistic metrics such as Bilingual Evaluation Understudy (BLEU) and F1, while overlooking broader learner-centered outcomes including cognitive load, retention, and smooth difficulty transitions [16, 17].

These challenges demonstrate a disconnect between linguistic fluency and pedagogical adaptivity. To address this, it is essential to establish a comprehensive framework that combines the mounting body of literature regarding factual grounding, psychometric rigor, and reinforcement-based adaptivity.

The significance of the study can be identified not only for the integration of existing modules, but also for theoretical as well as empirical contributions to adaptive testing and AI-question creation. PBAT offers a comprehensive assessment system that merges RAG-supported factual grounding, IRT-supported psychometric grounding, and a reinforcement-enabled adaptive delivery system using a restless Multi-Armed Bank (MARB) model, aligned with an assessment scheme for cognitive development according to Bloom's taxonomy of cognitive development and learning [18]. Moreover, the MARB-enabled cognitive profiling system not only measures cognitive growth through an actual continuous process, but is also capable of mapping the knowledge acquisition as well as the loss incurred even when a person is not exposed to a particular type of cognitive study, adding a significant dimension to the existing approach using a conventional adaptive learning process. Furthermore, PBAT offers a taxonomy-enabled filtering and calibration system that is adept at LLM-supported question generation aligned with cognitive development, eliminating redundancy in cognitive coherence. Finally, a multi-dimensional evaluation methodology for a comprehensive assessment approach for language quality, psychological validity, cognitive gains, and cognitive load can be identified from the study for a future research pathway for adaptive testing systems. Thus, in conclusion, PBAT connects the best adaptive testing with the intelligence of current LLMs to provide an intelligent solution for next-generation AI-enabled learning systems.

The rest of the paper is organized as follows: Section II reviews the existing literature on automated quiz generation and adaptive learning. Section III presents the proposed PBAT framework and its modeling. Section IV describes the dataset, experimental setup, and performance evaluation. Section V discusses the results and key findings. Finally, Section VI concludes the paper and outlines directions for future research.

II. Literature Review

The transition from template-based too deep learning models started with who utilized a sequence-to-sequence Long Short-Term Memory (LSTM) architecture for neural question generation using sequence-to-sequence LSTMs. The application of deep learning to this area resulted in a greater fluency of questions compared with template-based approaches; however, these systems do not allow for much control over the relevance or factual accuracy of the questions were generated. The introduction of attention and copy mechanisms in [19] improved on this earlier work by allowing the developers to keep the context more intact and thus reducing some of the hallucination problems from before. There have also been attempts to develop question generation systems that will allow users to control both the topic and the level of difficulty for the generated questions within LLM-based question generation systems [20]. Research on Quality-aware BLEU (Q-BLEU) was undertaken by Scialom *et al.* [21] to measure how well the questions generated by automated systems would provide an effective learning experience. Q-BLEU was beneficial for comparative purposes but did not cover adaptivity or providing learners with real-time feedback on their performance. The introduction and growth in popularity of generative AI has led many researchers to use a large number of pre-trained transformer models in their studies. introduced Bidirectional Encoder Representations from Transformers (BERT) and T5 into the research space as alternatives to the Traditional Question Generation (TGQ) model, resulting in improved semantic consistency between the generated questions and the input questions, as well as better grammatical fluency of the generated questions. While working on this project, developed grammar-based neural networks based on syntactic and part-specific features that improved the grammatical accuracy of the generated questions. However, the models developed by Hou did not consider the need for user adaptability. Research conducted by Khosravi *et al.* [2] demonstrated that as IRT continues to rise in popularity, IRT is superior to Classical Test Theory (CTT) in its ability to accurately assess the ability of a student using the quiz, the discrimination parameter, and the guessing probability associated with a quiz. reported the results of their research on the creation of adaptive assessments with the use of machine learning and psychometric models (e.g., deep RL). Through its evolution, RL methodologies have progressed in the area of learner-centered adaptive learning technologies. An example of RL-based quiz-style question selections that are informed by students' responses, creating maximum learning opportunities for all students in limited time frames. Using a latent variable sampling technique. Wang *et al.* [22] proposed a method to improve the diversity of questions presented to learners, although it did not establish a relationship between the questions and curriculum objectives. the RL for Question Generation (RL4G) framework for question generation and associated it with the development of high-quality questions that incorporate

difficulty levels and the distribution of reward signals; however, the learner's approach did not utilize objective measures of students' learning processes to calibrate the question quality. Malrick and Corbette [23] used the Generative Pre-trained Transformer (GPT) approach of prompt engineering in the author's study to achieve domain-specificity in question generation and control over format and content; however, again, there was no means to adjust either based on learners' experiences. An Educational Question Generation framework (EduQG), which employs Bidirectional and Auto-Regressive Transformers (BART) and T5's transformational architecture to generate questions in alignment with curricular guidelines. The EduQG framework provided good performance concerning linguistic quality and coverage but did not integrate mechanisms for the adaptivity expected in a learner-centered adaptive learning technology or provide feedback mechanisms for the learner. to generate embeddings for the purpose of continuously modelling learners' states. However, Zheng *et al.* [24] depended on synthetic data to conduct as per the author's investigation rather than collecting and analyzing information about real learners interacting with real content. Developed methods to modify LLMs for use in tutoring, where the LLMs iterate generating clarification questions within dialogues, thus providing some interactivity; however, this means that the models are not yet able to support strictly structured quiz-based interactions. Recently, the analysis of GPT-4's abilities in the area of educational applications has been performed. The authors found significant gaps between the factual accuracy of GPT-4 and how well its curriculum matches the requirements for education. One finding was that the authors believed the use of RAG could reduce hallucinations. A previous work by Daomin [25] proposed the use of multi-agent hybrid systems comprised of RL and LLMs to implement a dynamic assessment process; however, the complexity of these systems limited scalability to large classrooms. In contrast to these studies, there are emerging resources on LLMs and their integration with Vector Databases (VecDB). According to recent reviews, LLMs can use VecDB in order to quickly perform similar searches across variable-sized embeddings The use of VecDB also provides multiple levels of context support for the retrieval of information stored on LLMs and enables LLMs to recover factual correctness and reduce hallucinations and RAG overhead. How LLMs can beneficially use VecDB for question generation in RAG, alongside work by Lohr *et al.* [26] on RAG-based generation of semantically annotated learning objects and Aigerim *et al.* [27] on a low-resource approach to question answering, illustrates the advantages of the combination of RAG and LLM in generating content for education. A review of the literature indicates that NLP has progressed significantly from rule-based and surface-level techniques to more advanced methods driven by deep learning algorithms (neural networks), RL, and Large-scale Language Models (LLMs). Many limitations still exist within the areas of real-time adaptivity, learner modeling, reducing the frequency of hallucination, and ensuring all content produced aligns with educational objectives (pedagogical goals). Thus, our purpose in developing PBAT was to incorporate the benefits of both RAG and profile-aware sequencing into a single process so that our assessments are tailored to individual learners, reliable, and aligned with curricular objectives.

III. Methodology

In this section, we discuss the architecture of the proposed system, explore how its different parts work together, and share the key algorithms that make it function effectively. Fig. 1 depicts the architecture of the proposed system. It is a modular end-to-end intelligent assessment pipeline that automates the generation and adaptive delivery of quizzes. It is divided into three key modules: the Document Processing Module, the Quiz Generator Module, and the PBAT. Each module contributes parameters that flow through the system to enable contextual, personalized, and data-driven assessments.

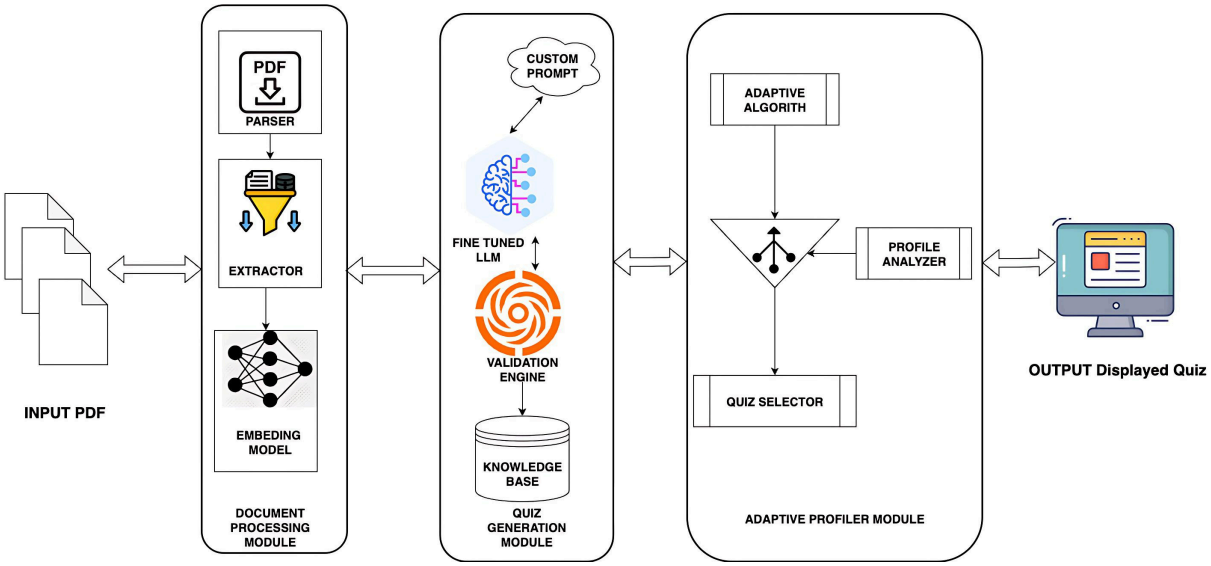


Fig. 1. PBAT system architecture.

A. Document Processing Module

This module marks the beginning of the pipeline chain and is intended for the preparation of raw instructional content for subsequent processing. Therefore, facilitating the ability to

translate the uploaded study resources, which are generally in the form of Portable Document Format (PDF) files, into a structured form that is adequate for subsequent process. The module had a four-stage pipeline. Initially, the PDF parser identifies the content of the document through.

The system comprises three main modules: a Document Processing Module for the extraction and embedding of information from incoming PDF files, a Quiz Generation Module for the production of quizzes through a fine-tuned LLM and a validation engine from a knowledge base, and an Adaptive Profiler Module for the selection of quizzes through an adaptive algorithm according to a user's profile, consisting of structural components such as titles, subtitles, tables, and text bodies. This parsing process establishes the structure and order of the original content. Then, a Content Extractor using a chunking algorithm breaks down raw text into meaningful units or 'segments' for various topics or ideas. Chunk-based representation helps ensure consistency within the context for question generation. In this step, let the knowledge corpus be a set of PDF files and research documents, represented by Eq. (1).

$$D = \{d_1, d_2, ..., d_n\} \quad (1)$$

Each document d_i undergoes parsing, preprocessing (e.g., stop word removal, sentence segmentation), and context extraction. The resultant content was segmented into semantic units. which is denoted as Eq. (2).

$$K = \{k_1, k_2, ..., k_m\} \quad (2)$$

After segmentation, the embedding model maps the units of text into numerical vectors of high dimensions. These embeddings preserve the semantic depth of the content without loss of information, allowing for more contextual comprehension. Within this framework, each unit k_i is transformed into a vector representation using the document embedding function *f,* which can be represented by Eq. (3).

$$V = f(K) = \{v_1, v_2, ..., v_m\}, v_i \in \mathrm{R}^d \quad (3)$$

The last step generates vector embeddings that represent the meaning of each segment in a retrieval- and computation-friendly format. These vectors are memorized for future use in the quiz generation module, where they are used as contextual anchors to ensure that the generated questions are accurate in terms of facts and directly relevant to the given material. Therefore, these vectors form the indexed knowledge base for RAG in the subsequent modules.

B. Quiz Generator Module

The second stage is where the process of quiz question generation from the knowledge base occurs. After analyzing and semantically labeling the study content, quiz questions are developed by employing world-class LLMs. The quiz question generation component initiates this process with the help of an application-predefined prompt where core output format prescriptions are given to the LLM regarding generating multiple-choice as well as true/false questions as per certain teaching requirements regarding clarity, level of difficulty, and applicability. The world-class fine-tuned LLM, coupled with the earlier developed contextual embeddings, starts developing questions syntactically as well as semantically precise.

To guarantee the accuracy and dependability of the questions produced, the Validator Engine conducted a thorough post-generation validation. It consists of relevance verification, duplication detection, and format verification. Only those questions that successfully passed this validation procedure were saved. Therefore, for a given learner, profile vector P and top-k content vectors are retrieved using a similarity search, denoted as Eq. (4).

$$C = R(P, V) = \{c_1, c_2, ..., c_k\} \quad (4)$$

These content chunks are then used to generate a set of candidate questions through a prompt-tuned or fine-tuned Large Language Model (LLM), denoted by Q_c = LLM(C). Then, the raw questions Q_c are validated using a combination of pedagogical rule filters and cross-referencing with a database of previously curated questions Q_{db}, producing a final filtered set represented as Q_f = F (Q_c, Q_{db}). At this stage, each question $q_i \in Q_f$ is annotated with IRT parameters as IRT $(q_i) = (a_i, b_i, c_i) \in \mathrm{R}^3$, where a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the guessing parameter. Finally, the probability of a correct response was modeled using Eq. (5).

$$P(\theta) = c + \frac{1-c}{1+exp(-\alpha(\theta-b))} \quad (5)$$

The final author-approved questions are stored in the Question Bank (or Item Bank), a centralized database labelled with rich metadata, such as difficulty level, question type, and linking topic or chapter. This organized labeling allows the system to automatically retrieve and tailor questions to various learners during adaptive assessments.

A. Adaptive Profiler Module

The third and most important module is called PBAT. It uses an adaptive algorithm to create a personalized assessment based on learners' ever-evolving ability profiles. When a learner takes a PBAT, it begins with the Question Selector, which typically begins the quiz with a medium-level question. After that, using adaptive logic, subsequent questions will change in difficulty based on the learners' responses to the previous question. The front end of the PBAT provides a quiz interface that displays all the questions to the learner and tracks the answers as well as other information, such as how long the learner takes on each question. This module has two specific sub-algorithms:

(1) **Adaptive Algorithm**—Responsible for estimating the learner's ability dynamically (θ) and changing the difficulty of questions based on real-time data.
(2) **Profile Analyzer**—Uses a MARB model to track learner performance across different topics and levels of difficulty.

Both algorithms were developed using five foundational CAT Principles: calibrated item bank, starting rule, item selection rule, scoring rule, and stopping rule. All these principles provide an efficient data-driven and learner-specific assessment process.

B. Adaptive Algorithm Using RAG and Pedagogical Rules

The adaptive algorithm generates quiz questions that are customized to the learner's profile and topic. It begins by deploying the RAG paradigm, which incorporates dense retrieval of relevant context with the generative capabilities of pretrained language models. This module generates quiz questions tailored to the learner's profile. It starts by retrieving the top-k relevant knowledge chunks from the indexed knowledge base using similarity-based retrieval on the learner profile vector P. These chunks, denoted by $C = \{c_1, c_2, ..., c_k\}$, are passed to a fine-tuned large-language model to generate candidate questions Q_c = LLM(C).

However, not all generated questions are pedagogically suitable. In turn, a rule-based filtering engine was applied. These rules are based on Bloom's taxonomy, concept coverage, repetition control, novelty scoring, and difficulty control. Questions are filtered for hallucinations using named entity checks and compared with source retrieval spans. Only those questions that met the difficulty and pedagogical criteria were forwarded for learner-specific ranking and adaptation. This ensures that the input chunks are relevant and factual, thereby mitigating the hallucination risks that downgrade the LLM's performance. Consequently, the adaptive algorithm is structured in the CAT framework, which consists of the following subcomponents.

1) Calibrated item bank

The calibrated item bank was utilized as the primary dataset in the adaptive testing engine. The item bank now contains psychometric parameters that define each item according to the IRT model. The psychometric parameters for the item bank included the b (difficulty), a (discrimination), and c (guessing) parameters. The psychometric parameters were estimated through statistical calibration of the test pilot data via logistic regression. To determine the probability that a student with ability θ can respond correctly to an item, Eq. (5).

Hence, the psychometric parameters allow the system to provide continuous estimates of the probability of success for any item that can be chosen in real time. This is possible because the psychometric parameters of each item in the item bank are calibrated to allow for the adaptive selection of items to achieve maximum measurement accuracy and consistency, regardless of the learner's ability level.

2) Starting rule

Prior to an adaptive testing interaction, the learner's initial proficiency score θ_0 was established by the starting rule. In CAT systems, the preferred method to calculate this initial proficiency score has usually been by using a standardized Z-score, where the average score is represented by $\theta_0 = 0.0$. Conversely, this study utilizes a more detailed approach through the integration of available learner metadata, such as previous performance data, prerequisite test results, and confidence data, on a domain level. This customization of the Starting Rule allows the selection of initial questions to be based on greater confidence levels while minimizing the uncertainty associated with a learner being cold-started.

The starting rule can also be enhanced to include advanced methodologies in which both expert opinion and population statistics are incorporated to enhance the convergence speed of the assessment process during the early testing stage. The Starting Rule provides the learner with its initial assessment ability estimate θ_0 and supports the learner in progressing through the assessment process adaptively. θ_0 is typically assigned a fixed value of $\theta_0 = 0.0$), but this approach incorporates a more advanced and personalized method of estimating the learner's proficiency level through the use of a Bayesian Prior Estimation [28] that sets θ_0 by means of Eq. (6).

$$\theta_0 \sim \mathrm{N}(\mu, \sigma^2) \tag{6}$$

where μ represents the prior estimate of the learner's ability, and σ^2 reflects uncertainty about that estimate. If no learner-specific data are available, $\mu = 0$ and $\sigma^2 = 1$ are used, assuming average ability with moderate uncertainty. When prior data are available, Grade Point Average (GPA) diagnostic test scores, μ, *are* inferred using a regression-based prediction model using Eq. (7).

$$\mu = \beta_0 + \beta_1 \cdot \mathrm{GPA} + \beta_2 \cdot \mathrm{PS} + \beta_3 \cdot \mathrm{SR} \tag{7}$$

where PS is the previous quiz score and SR is the self-rating. The aforementioned initialization approach also acts as a remedy for the cold start problem, establishing a correlation between the first chosen question and the supposed level of aptness of the learner. This scoring technique is appropriate for the Bayesian models: Maximum A Posteriori (MAP) or Expected A Posteriori (EAP) method. This method selects the parameter that achieves the highest posterior probability using Eq. (8) below:

$$\theta_{MAP} = arg\ \max\nolimits_{\theta} (logL(\theta) + log\pi(\theta)) \tag{8}$$

While the EAP computes the expected ability across the posterior distribution using Eq. (9).

$$\theta_{EAP} = \frac{\int^{\theta} .L(\theta)\pi(\theta)d\theta}{\int^{L} .(\theta)\pi(\theta)d\theta} \tag{9}$$

These methods ensured consistent and personalized adaptation throughout the testing session. PBAT can be understood as a probabilistic adaptive assessment paradigm wherein evidence grounding, latent trait ability modeling, and sequential decision-making are integrated as interlocking parts of a single process, with RAG framing the question generation process over a grounded evidence space, thereby eliminating the possibility of hallucinations. IRT specifies learner ability mapping θ as a probabilistic function of item characteristics as inputs, with MARB implementing the policy as a reward-control process over an evolving knowledge state. Instead of being a bag of tricks, PBAT captures the integration of grounding, calibration, and policy as an interlocking set of parts within a unique paradigm of adaptive assessment.

It operates on a set of basic assumptions: that learner knowledge can be represented as a latent trait θ, estimated from item-response behavior; that knowledge states evolve over time through learning and forgetting, and can thus be updated via reinforcement-based interaction signals. The framework further assumes that question reliability improves when generation is constrained by retrieved evidence rather

than produced through unconstrained language-model sampling, and that assessment efficiency increases when item selection is formulated as a reward-maximization process rather than a fixed progression rule. Together, these assumptions provide the theoretical basis that links RAG, IRT, and MARB within a unified adaptive learning framework.

Although PBAT was instantiated in this study for quiz-based assessment, the framework generalizes naturally across a wide range of learning and assessment applications. Such an application targets settings where task content can be naturally rooted in retrievable evidence, where performance can be represented through latent-trait estimation, and where sequential interactions would most benefit from adaptive policy optimization. These same theoretical components of evidence-constrained generation, psychometric calibration, and reinforcement-based sequencing can be re-parameterized for various other domains, such as programming assessment, concept-diagnostic testing, and language-learning exercises, without changing the underlying framework. The complete adaptive algorithm is depicted in Algorithm 1.

Algorithm 1: Adaptive Question Generation Using RAG and Pedagogical Rules

Input: Learner Profile P, Knowledge Base K
Output: Filtered and calibrated question set Qf

1. Retrieve top-k relevant chunks: C ← Retrieve(P, K)
2. Generate candidate questions: Qc ← LLM(C)
3. Apply rule-based filters: Qf ← Filter(Qc)
4. Annotate each qi ∈ Qf with IRT parameters (a, b, c)
5. Return Qf

C. Profile Analyzer Using MARB

The MARB model selects questions while attempting to elicit the maximum amount of diagnostic information. The MARB model differs from classical multi-armed bandit models because the arm (representing a specific topic × difficulty pair) can adapt over time, even if it is not selected. This is an important pedagogical distinction, since students' degrees of understanding may improve or decline within a topic through exposure, retention, and time. In MARB, each arm has a belief state (belief about the learning state of the topic × level of difficulty) that evolves based on historical data (e.g., student response accuracy, response times, confidence ratings). MARB applies a state transition model to forecast the expected reward from a particular arm at this instant. These expected rewards can be expressed as modifications to learning, rates of improvement, or reinforcement values. In contrast to classical multi-armed bandits, wherein an arm is static until selected, the MARB model allows for the evolution of an arm depicting a concept × difficulty tuple, regardless of whether that arm has been selected. This feature is particularly useful in education, where students are likely to have multiple opportunities to get better or worse on a particular topic over time, based on the level of exposure to it. The profile analyzer is another component of the CAT framework, similar to the adaptive algorithm, and is referred to in further detail in this section.

1) Item selection rule

The Item Selection Rule determines the most informative question to pose to a student based on their estimated aptitude. A learner's aptitude represents the uncertainty in the IIT's estimation of the aptitude (θ) computed using Eq. (10).

$$I(\theta) = a^2 \cdot \frac{(1-P(\theta))}{P(\theta)} \tag{10}$$

The next item was chosen with the question that had the highest Fisher Information score. Consequently, learners will continue to face challenges at the edge of their capabilities. Therefore, the maximum diagnostic resolution and maximum learning gains can be achieved.

Alternative methods for selecting the next item include the Maximum Posterior Weighted Information Method (MPWI) and Kullback-Leibler divergence; however, Fisher Information remains the most efficient method for Item Selection and is the easiest to interpret.

2) Scoring rule

After each response, the learner's ability estimate θ needs to be updated. This is achieved using statistical estimation techniques:

- Maximum Likelihood Estimation (MLE) maximizes the likelihood of observed responses but is undefined when all responses are correct or incorrect.
- MAP addresses MLE's limitations by incorporating a prior distribution over t*heta*, computed using Eq. (11).

$$\theta\_\text{MAP} = \arg\max\_\theta (\log L(\theta) + \log \pi(\theta)) \tag{11}$$

The scoring rule ensures that the learner's trajectory through the quiz reflects the true learning evolution and informs the next selection.

3) Stopping rule

The Stopping Rule outlines the duration of a particular quiz session. With traditional assessments, the number of questions asked is normally fixed within that assessment, whereas CAT is a more rigorous and responsive stopping method than simply stopping when the number of questions reaches a designated amount, as is done with traditional assessments.

In a typical assessment, the Standard Error of Estimation (SEE), θ, would be calculated and when SEE drops below ε, the computerized adaptive assessment would cease, computed using Eq. (12).

$$SE(\theta) < \epsilon \tag{12}$$

The stopping criteria allow for the possibility of pseudo-answers being generated by an assessment, thus preventing a computerized assessment from becoming obsolete due to excessive duration that could create assessment fatigue in students who might otherwise wish to continue the quiz. However, to ensure that students finish with accurate assessments as quickly as possible, many computerized assessments implement maximum question limits and/or a time period associated with stopping the assessment. Adaptive stopping strategies enable students to receive a more individualized testing experience while providing optimized learning and evaluation capabilities. The complete profile analyzer algorithm can be found in Algorithm 2.

Algorithm 2: Profile-Based Question Selection using MARB

Require: Filtered questions Qf, Learner history H, Arm set A

1. Initialize ability $\theta 0 \sim N(\mu, \sigma^2)$
2. While Stopping Rule not satisfied, do
   a. Update belief state for each arm $a \in A$
   b. Estimate reward R(a)
   c. Select arm $a^* = \arg\max R(a)$
   d. Select question $q \in Qf$ matching a^*
   e. Present q and collect learner response
   f. Update θ
3. End while
4. Return adaptively generated quiz

4) Quiz selector

This is the final stage, where the outputs from the Adaptive Algorithm and the Profile Analyzer are combined to create a unique quiz for every learner. The resulting quiz will meet educational requirements, eliminate repeated assessments of similar concepts, and provide a range of cognitive levels and all levels of difficulty.

The Profile Analyzer identifies questions across all content areas and captures a wide variety of Bloom's levels and concepts for each learner. Accordingly, it reduces redundancy by not using a concept that has already been evaluated with a high degree of certainty; it prioritizes questions that have the greatest amount of Fisher Information related to the learner's level of ability (θ); it also enforces stopping rules that establish either a maximum number of questions or an acceptable level of confidence before a learner is considered proficient. Additionally, candidate questions are reordered continuously in real time to maintain logical sequencing, provide the current level of challenge, and maintain learner immersion.

Finally, the corresponding quiz is delivered through an interactive user interface. This represents the final stage of quiz development. The interface is designed to promote clarity and enjoyment while providing dynamic difficulty adjustments based on learner performance. Additionally, the learner's responses to the questions are logged into the background, allowing the system to improve its performance in future assessments. Overall, the displayed quiz is designed to provide an accurate diagnosis of the learner's abilities while also providing an adaptive and personalized learning experience. Algorithm 3 presents the complete working of the adaptive quiz generation.

Algorithm 3: Integrated Adaptive Quiz Assembly

Require: Learner profile P, Knowledge base K, History H, Arm set A

1. Retrieve top-k chunks: C ← Retrieve(P, K)
2. Generate question set: Qc ← LLM(C)
3. Apply pedagogical filters: Qf ← Filter(Qc)
4. Annotate Qf with IRT parameters (a, b, c)
5. Initialize $\theta_0 \sim N(\mu, \sigma^2)$
6. While Stopping Rule not satisfied, do
   a. Update belief state for each arm $a \in A$
   b. Compute reward R(a)
   c. Select arm $a^* \leftarrow \arg\max R(a)$
   d. Select question $q \in Qf$ from arm a^*
   e. Present q and collect learner response
   f. Update θ using MAP or EAP
7. End while
8. Finalize quiz with balanced coverage and logical sequencing
9. Return adaptively generated quiz

PBAT has split the workflow into offline and online components, giving the system the ability to operate well in limited-resource educational environments. The compute-heavy tasks of generating RAG-based questions and estimating IRT parameters are completed offline, once per course. In our implementation, approximately 200–300 questions were generated within 35–45 min using a single consumer-level graphics card (NVIDIA RTX 3060 / 12GB). The newly calibrated question bank created monthly is available to the course owner at an earlier time.

When using PBAT for online quizzes, only lightweight computations are necessary to estimate students' abilities, update multiple assessments, and load student answer vectors. Standard CPU servers complete the entire online quiz within an average of 78–110 ms per question selected; therefore, there is no computation necessary during student use. This method substantially reduces hardware requirements and enables institutions that do not have the resources for advanced hardware to deploy PBAT.

IV. Experimental Setup and Results

In this section, the methods used to create the dataset, implement the experiments, baseline models, metrics used to evaluate the results of the PBAT framework, and data analysis of the results are presented. All experiments were conducted in a controlled environment to provide an equitable basis for comparisons across multiple types of PBAT configurations. Additionally, as part of evaluating the scalability of PBAT, benchmark tests were performed utilizing typical computers. All offline generation tests were performed on an NVIDIA RTX 3060 GPU, whereas all online adaptive tests used a CPU with Intel i7 without a GPU. Both the MARB selector and the IRT Updater scripts operated for less than 0.12 seconds per item, confirming that PBAT components may be installed in both school and professional settings that use standard PC or laptop configurations

A. Baseline Models

For benchmarking, PBAT was compared with three established techniques:

- **3-Parameter Item Response Theory model (IRT-3P)**: IRT with a 3-Parameter Logistic model, which selects and orders questions based on item difficulty, discrimination, and guessing probability.
- **Non-Adaptive Static Quiz**: A fixed LLM-driven model that generates questions without personalization, learner modeling, or adaptive sequencing.
- **RL-Adapt**: An RL-based adaptive generator that optimizes a policy using a composite reward function capturing linguistic quality, question answerability, and learner alignment.

These baselines were selected to test the PBAT against psychometric, static, and RL-based adaptive strategies.

B. Dataset

In this section, we describe the datasets that we used and our experimental approach for evaluating the PBAT Algorithm proposed in this study. We used a set of course materials from standard university-level courses in Information Technology Engineering to create an instance of pedagogical relevance and domain alignment. We included

four core subjects: Cloud Computing, Machine Learning, Artificial Intelligence and Cybersecurity. We generated approximately 30 to 40 questions per subject using LLMs that we guided using structured prompts designed specifically for each subject area.

The process we developed for generating the questions was based on a set of carefully developed prompts that we created to reflect the different levels of Bloom's Taxonomy from the Knowledge Recall level to the evaluation-based tasks. We included representative examples of the prompts in Appendix A to provide a level of transparency and reproducibility of this methodology. The prompts were developed to guide the LLM to produce curriculum-aligned and semantically rich questions while avoiding the production of overly generic outputs.

Nooralahzadeh *et al.* [19] proposed an approach to refine the generated output using reinforcement filters. The approach included defining and implementing a reward function used to rate the generated questions along three dimensions: (1) semantic relevance to the content of the source material based on the cosine similarity of question embeddings, (2) the degree of linguistic fluency of questions based on grammatical and syntactical standards, and (3) the degree of match between the generated questions and Bloom's taxonomy of cognitive learning objectives, as measured by a taxonomy-aware classifier. Only questions that achieved a high score in each dimension were included in the final evaluation dataset to generate the quiz questions. This dataset is designed to provide an even representation of Bloom's taxonomy levels and to have sufficient coverage of the subject area in question, based on domain-specific vocabulary. The linguistic analysis of the dataset demonstrated that there is a high degree of domain-specific vocabulary as well as consistency of use, making the data appropriate for conducting adaptive learning research. Additionally, a simulated performance dataset of 100 students at three skill levels (low, medium, and high) taking quizzes designed by PBAT using IRT-3P and conventional methods was generated (see Table 1).

Table 1. Student performance dataset parameters

Parameter	Description
Student ID	Unique identifier for each student
Ability Level	Categorized as Low, Medium, or High based on initial profile
Quiz Technique	Quiz mode used (PBAT, IRT3P, No Adapt)
Attempt No	1st, 2nd, or 3rd quiz attempt
Time Taken (s)	Time taken to complete each quiz
Questions Attempted	Number of questions answered per quiz
Average Score	Final score normalized to 100
Feedback Score	Learner's feedback on a Likert scale (1 to 5)

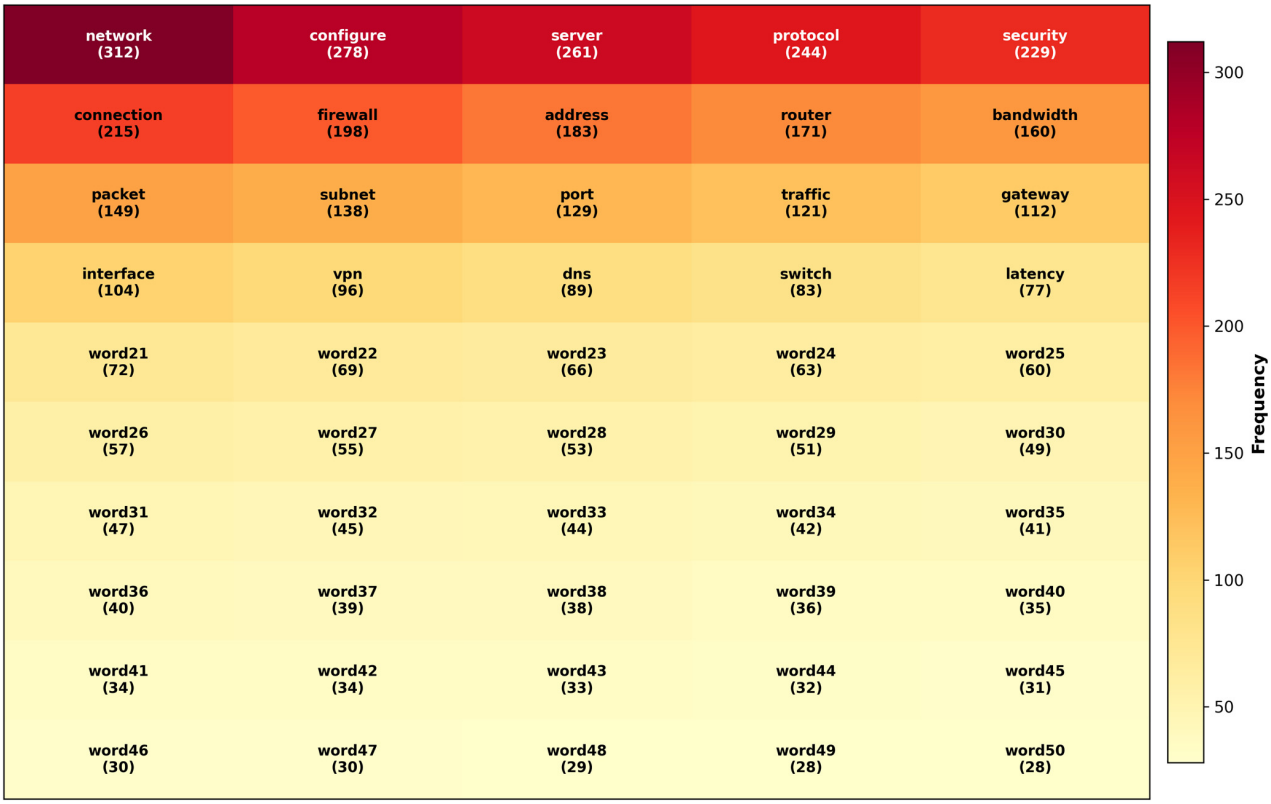


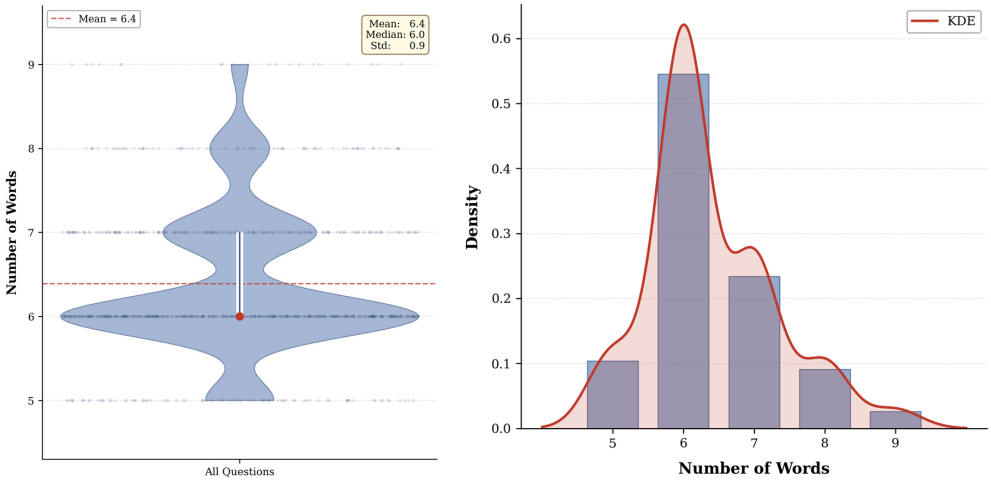
Fig. 2. Heat map visualization of the 50 most frequent words across the generated dataset, highlighting their frequency intensity and contextual prominence.

To support reproducibility and open science, the complete dataset, implementation code, and experiment scripts are publicly available at https://github.com/PanksMish/pbat-framework.git. As this study focused on system-level evaluation, all learner responses were simulated using predefined ability distributions and IRT-based response models; no real student data were used in this phase.

Fig. 2 illustrates that the high-frequency clusters correspond to domain-specific keywords such as "cloud," "model," "data," and "learning," confirming a strong curriculum alignment. The consistent color gradients across topics indicate balanced lexical diversity, ensuring that the generated questions remain contextually grounded, varied, and pedagogically relevant.

Furthermore, Fig. 3 illustrates that the most frequently generated questions were concise, with the majority falling

between 8 and 12 words. The narrow density curve indicates linguistic consistency and controlled verbosity, suggesting that the question generation process-maintained brevity while Ensuring semantic clarity—an essential characteristic for adaptive assessments where comprehension speed and cognitive load need to be balanced.



(a) (b)

Fig. 3. Distribution of question lengths in the generated dataset.

C. Experimental Protocol

Each learner group was exposed to quizzes generated by the PBAT, IRT-3P, RL-Adapt, and Non-Adaptive methods. The evaluation considered both the following:

- **Linguistic Metrics**: BLEU-1, BLEU-4, F1-Scores, perplexity, and vocabulary diversity.
- **User-Centric Metrics**: Hallucination rate, retention after 48 h, cognitive load measured using the National Aeronautics and Space Administration Task Load Index (NASA-TLX) learner feedback (Likert-scale), and difficulty transition stability. The results were averaged over 10 randomized trials. Statistical significance of the observed differences was validated using one-way Analysis of Variance (ANOVA) followed by Tukey's post-hoc test at $p < 0.05$. Table 2 lists values of all the hyperparameters.

Table 2. Hyperparameter settings used for PBAT and baseline models. These values were optimized via grid search on a held-out validation split

Parameter	Value
LLM Backbone	GPT-3.5 (OpenAI)
Max Question Length	20 tokens
Fine-tuning Epochs	5
Batch Size	16
Learning Rate	2×10^{-5}
RL Algorithm	REINFORCE with Baseline
Discount Factor (γ)	0.95
Exploration Policy	ϵ-greedy ($\epsilon = 0.1$)
Stopping Rule	Max 15 questions or ± 0.05 change in θ
Reward Weights	Relevance: 0.4, Fluency: 0.3, Bloom-alignment: 0.3

D. Results on Linguistic Metrics

Fig. 4 presents a comprehensive comparison of the PBAT, RL-Adapt, IRT-3P, and non-adaptive approaches across multiple evaluation metrics in Fig. 4(a), BLEU-1 and BLEU-4 scores show that PBAT and RL-Adapt achieve both higher medians and narrower spreads, suggesting that they consistently generate syntactically correct and contextually aligned questions. The BLEU-4 improvements are particularly notable for PBAT, reflecting its ability to preserve longer phrase structures and capture deeper semantic matches. Fig. 4(b) highlights the F1-Scores, where PBAT demonstrates the most balanced and stable precision–recall relationship, ensuring that generated questions align closely with the target answers. RL-Adapt perform competitively but with less stability, while the non-adaptive approach exhibits high variability, indicating inconsistent learner alignment. In Fig. 4(c), the perplexity results reveal that PBAT and RL-Adapt maintain lower values, producing grammatically fluent and natural-sounding questions that enhance learner comprehension. Unlike earlier work, Static and IRT-3P models display elevated perplexity, often leading to awkward phrasing that may hinder engagement. Fig. 4(d) illustrates vocabulary diversity, with PBAT and RL-Adapt covering a broader lexical range, thereby reducing redundancy and fostering conceptual breadth while avoiding rote memorization. Finally, Fig. 4(e) shows the F1-Score improvement after fine-tuning, where PBAT exhibits the highest gain, underscoring its adaptability to domain-specific content and strong potential for continual performance enhancement. RL-Adapt benefits moderately, whereas Static and IRT-3P show minimal improvement.

E. Results on User-Centric Metrics

As illustrated in Fig. 5(a), PBAT's hallucination rate of 8.2 % is the lowest of all the models tested and confirms that RAG-based validation is effective in filtering out incorrect and irrelevant questions. PBAT's improvement in retention performance, as illustrated in Fig. 5(b), shows that PBAT allows students 15.6% better recall of knowledge after seven days compared to static and demonstrates the pedagogical effectiveness of PBAT. Fig. 5(c) displays the NASA TLX cognitive load scores, which indicate that the PBAT has the least cognitive load, thereby providing the best balance between challenge and clarity.

As seen in Fig. 5(d), students' ratings of their experiences with PBAT are the highest of all models. Fig. 5(e) provides a graphical representation of the stability of the transition in difficulty for each evaluated learner type. The transition curve of PBAT for learners appears to be smoother over time in real

time, indicating PBAT's ability to adapt and present questions with difficulty that evolve along with learner performance, rather than create sudden shifts in question difficulty, which could have the opposite effect of maintaining student engagement.

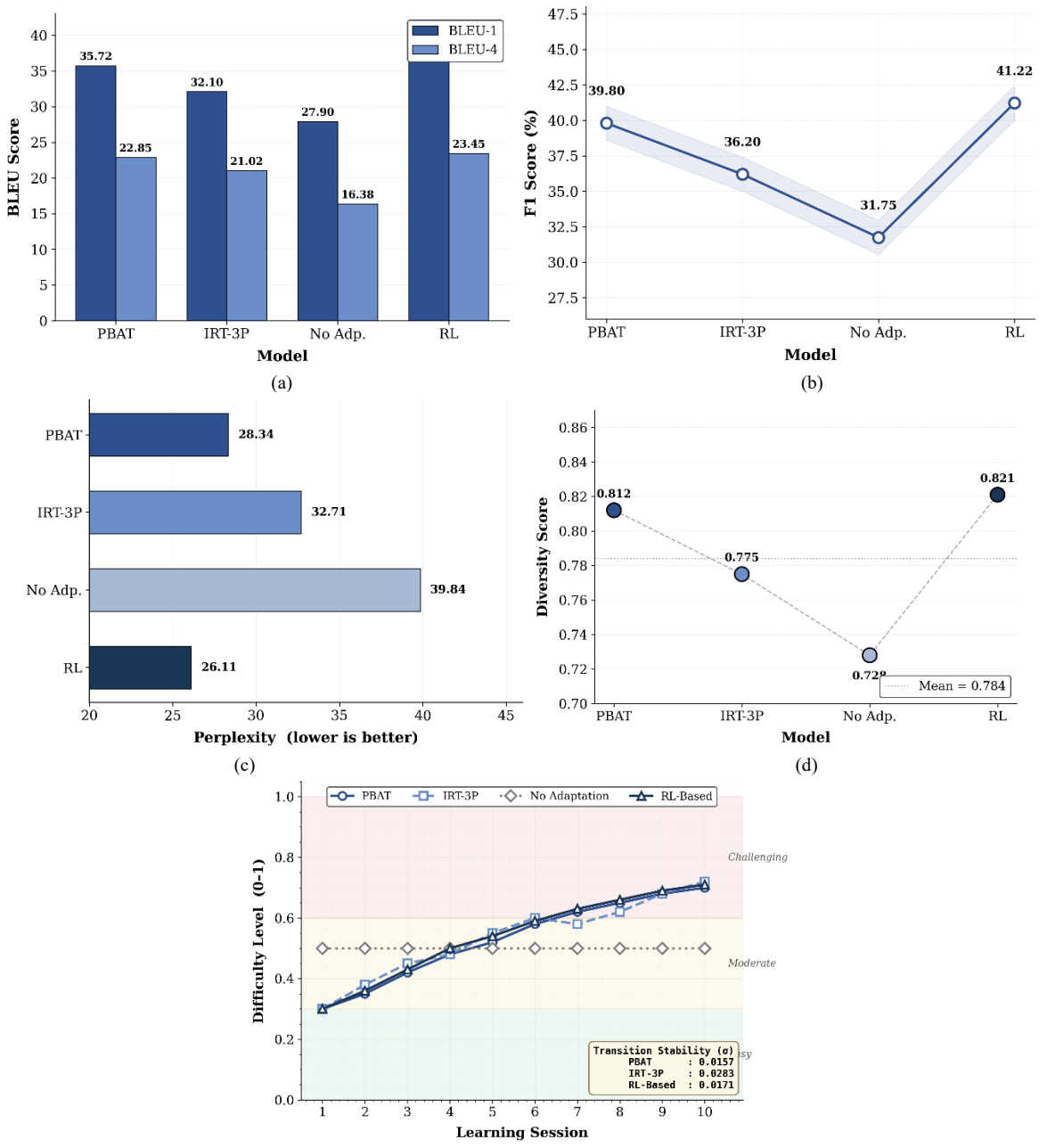
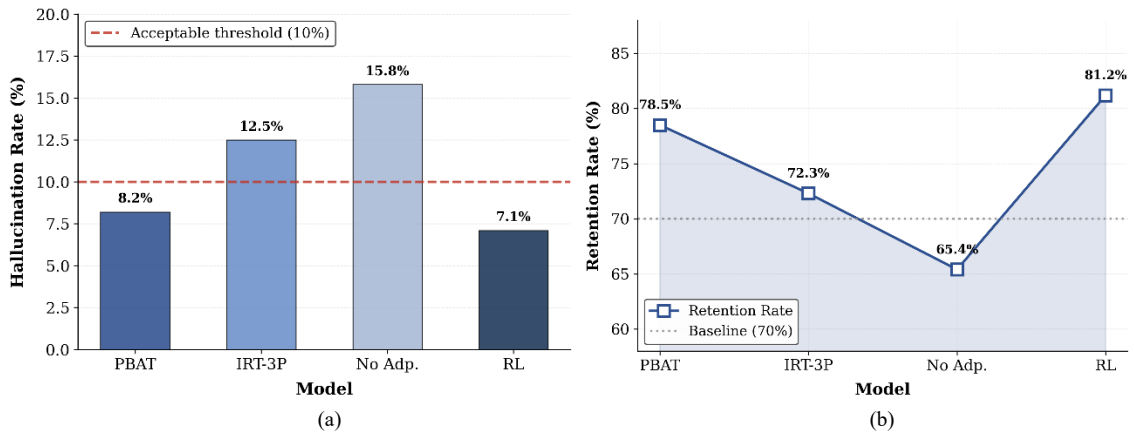


Fig. 4. Linguistic performance metrics for PBAT and baselines (a) BLEU-1 & BLEU-4, (b) F1-Score, (c) perplexity, (d) vocabulary diversity, and (e) fine-tuning Gain. PBAT consistently achieves higher BLEU/F1, lower perplexity, and broader lexical coverage than RL-Adapt, IRT-3P, and non-adaptive models, confirming superior linguistic fluency and factual consistency.



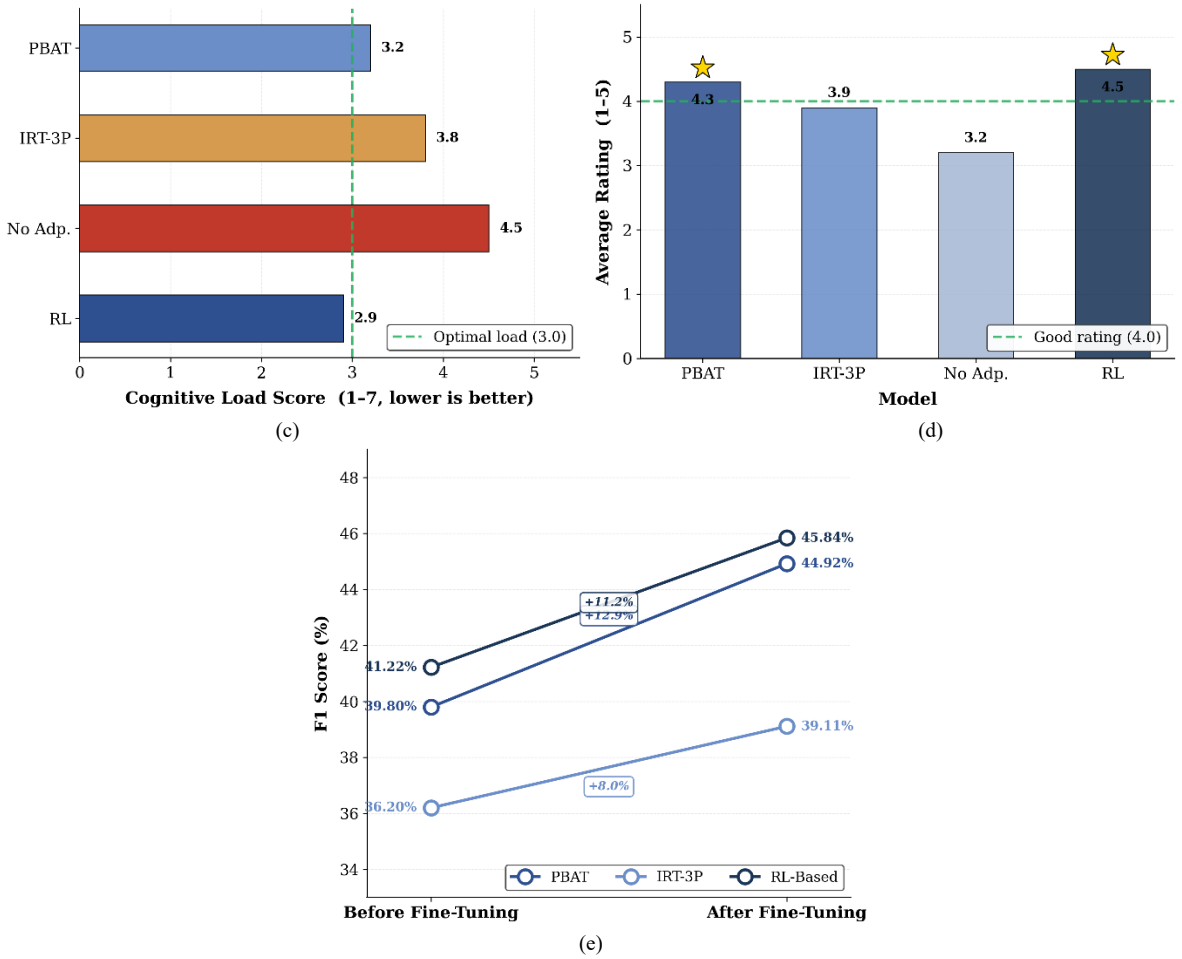


Fig. 5. User-centric evaluation metrics for Profile-Based Adaptive Testing (PBAT) and baselines (a) hallucination rate, (b) retention after 7 days, (c) cognitive load (NASA-TLX), (d) learner satisfaction, and (e) difficulty Transition Smoothness. PBAT demonstrates reduced hallucination, improved retention, lower cognitive load, and smoother difficulty progression, validating its adaptivity and learner-centered performance.

PBAT's implementation, along with associated datasets and experiments are available at: https://github.com/your-repo-link for open access to facilitate reproducibility and continued investigation into the methodology employed in the study. During a follow-up manual review of the remaining 8.2% of hallucinated items, three major themes emerged. Each theme accounted for an approximate percentage of the total number of hallucinated items. Misinformation is defined as factual errors in the definition, entity name, or quantity produced by the LLM (approximately 60% of all cases). Logical inconsistencies occurred primarily via the presentation of ambiguous keywords, conflicts between answer choices, and incongruity between the textual leading phrase/appropriate keyword(s) and correct answer choices (about 25% of cases). Pedagogical Drift is a term used to describe cases where the learning objective did not reflect the appropriate Bloom's Taxonomy (approximately 15% of cases). This provides another avenue for classification that aids in a better understanding of the remaining hallucinations. Most hallucinations were mild factual inaccuracies rather than severe contradictions. The Bloom-level filtering and RAG grounding eliminated high-impact errors, leaving only low-severity cases that are amenable in future work through deeper semantic consistency checks.

V. Discussion

Both linguistic and learner-centered evaluations proved that, in all respects, PBAT outperforms traditional models using IRT-3P and psychometric measures when generating questions with respect to quality, adaptability, and usability. By utilizing RAG, Bootstrapping using IRT, and MARB to generate questions through an iterative process, PBAT generates contextually relevant, pedagogically unifying, and user-centric question types.

In terms of Linguistic Quality, PBAT has superior BLEU scores and lower perplexity than IRT-3P and static models due to fewer hallucinations and improved semantic relevancy between questions generated by PBAT and the content covered within the curricular framework. In addition to being domain-specific, questions generated by PBAT have improved pedagogical value because smooth transitions are maintained in terms of the level of difficulty, which mitigates overwhelming feelings among learners. Additionally, IRT calibration and MARB selection, used as tools for determining question difficulty, allowed PBAT to avoid abrupt transitions from one question complexity level to the next and equally keep learners engaged regardless of the learner profile category.

Furthermore, adaptability through PBAT has been further validated by means of lower Cognitive Load on the NASA-TLX and greater Learning Retention rates at 7 days post-testing. The time frame indicates that PBAT successfully achieves a balance between challenging learners and, at the same time, ensures understanding.

The user experience associated with PBAT improved substantially. The PBAT was rated the highest for user

satisfaction among the instructional tools evaluated, and users found the quizzes to have an increased sense of being personalized, experienced as more closely reflecting users' previous knowledge, and feeling more natural. Statistically significant testing indicated that PBAT's increases on BLEU-4, F1, and retention were significant ($p < 0.05$). The results of ablation studies from PBAT clearly demonstrate that PBAT's significant benefits come from having multiple modules that complement one another, rather than from LLM outputs. In conclusion, PBAT has significant potential for adaptive and scalable educational assessment in the real world.

VI. Conclusion

This study addressed the two major challenges of AI-generated quizzes—(1) hallucinations associated with generated questions and (2) an insufficient degree of meaningful personalization—by employing a RAG process of fact-checking in conjunction with PBAT. Therefore, the quizzes generated by the system will be accurate and context-based and will adapt in real time the proposed level of difficulty depending on an individual learner's success or lack of success during the quiz. A series of experiments were conducted that evaluated the outputs using the metrics of BLEU, F1, perplexity, cognitive load, vocabulary diversity, and learner feedback. The results showed that the PBAT produced significantly better results based on each of these evaluations when compared to both an IRT-3P model and a static approach to quiz generation. Overall, PBAT had the highest BLEU-4 and F1-Scores, the lowest level of cognitive load, and the greatest degree of vocabulary diversity, while achieving a consistently stable development of levels of difficulty over time, thereby allowing for learner engagement without overload. Learners gave consistently higher ratings for PBAT than for IRT-3P and the static approach, demonstrating that PBAT was capable of producing relevant and motivating assessments for all learners. While IRT-3P produced credible outputs, it did not have the capability of adjusting dynamically during the process. Conversely, the static approach created outputs that were inflexible and, therefore, offered limited opportunities for heightened engagement and interest. Through the employment of RAG to reduce hallucinations combined with PBAT's Adaptive Profiling, the system generates quizzes that are factually accurate and, at the same time, generates customized quizzes based on the ability levels of individual learners. The result is a learning experience that is both highly effective and highly engaging, thereby resulting in greater understanding and long-term motivation towards further academic achievement.

While PBAT seems to work well with structured English materials, there may be differences in how it behaves when interacting with multilingual databases or noisy or highly specialized domain-based information. PBAT evaluation was not conducted using cross-lingual prompts or domain-specific terminologies, which are associated with higher levels of hallucination. Therefore, extending PBAT to accommodate multilingual datasets and complex domain materials will be a principal focus of future research. The PBAT is required to be validated through actual student test results to determine if it accurately measures the learning that takes place on a day-to-day basis and whether PBAT is effectively utilized with multiple formats of content (i.e., traditional textbooks, digital textbooks, other materials). The ability to adapt PBAT to various venues and locations will allow for continuous improvement of question generation, and thus, an increase in the accuracy and fairness associated with questions produced by PBAT. PBAT's ability of PBAT to provide ongoing feedback through RL will result in continued improvements in both question generation and policy associated with the use of PBAT in practice. The commitment to continued efforts to reduce bias within PBAT will allow the company to establish PBAT as an ethical, equitable, and learner-driven AI assessment tool. As previously mentioned, PBAT is scheduled to be validated using real students to assess their actual learning behaviors; refine its MARB Adaptive Learning Model; evaluate long-term learning gain; and assess its fairness across diverse student populations in realistic classroom and lab environments.

Appendix A: Prompt Design and Rationale

As the proposed PBAT framework relies on the quality of the questions produced, we need prompts to effectively instruct the language model (LLM) to create questions that would be considered technically correct as well as pedagogically sound. To do this, we created prompts that specifically followed Bloom's taxonomy of educational objectives, used subject matter content from the same domain, and communicated to the LLM clearly and naturally. This process allows the dataset produced to be representative of the types of assessments students typically encounter in traditional educational settings.

A. How the Prompts Were Designed

Each prompt follows a simple but effective structure.

- **Instruction**: Clearly state the number and type of questions required (e.g., short-answer, scenario-based).
- **Cognitive Focus**: Include action verbs chosen from Bloom's taxonomy (e.g., "Define," "Apply," "Analyze," "Evaluate").
- **Domain Context**: Specify the subject and sub-topic (e.g., virtualization in cloud computing, supervised learning, intrusion detection systems).

B. Examples of Prompts

Here are some representative prompts and its sample outputs:

(1) **Knowledge Level (Cloud Computing)** Prompt: *"Generate 5 short-answer questions on virtualization in cloud computing. Use verbs such as 'Define', 'List', and 'Describe'."* Example Output: *"Define virtualization and explain its role in resource abstraction."*

(2) **Application Level (Machine Learning)** Prompt: *"Create 3 scenario-based questions on supervised learning. Use Bloom's verbs 'Apply', 'Demonstrate', and 'Use'."* Example Output: *"Given a dataset of house prices, apply a regression model to predict unseen values and justify your approach."*

(3) **Analysis Level (Artificial Intelligence)** Prompt: *"Generate 4 analytical questions on A* search algorithm. Use verbs like 'Compare', 'Differentiate', and 'Analyze'."* Example Output: *"Compare A* search with Dijkstra's algorithm in terms of optimality and time*

complexity."

(4) **Evaluation Level (Cybersecurity)** Prompt: *"Formulate 3 evaluation-based questions on intrusion detection systems. Use verbs such as 'Evaluate', 'Critique', and 'Assess'."* Example Output: *"Evaluate the effectiveness of anomaly-based intrusion detection systems in detecting zero-day attacks."*

C. Principles Behind Prompting

- **Coverage**: Every core subject (Cloud Computing, Machine Learning, Artificial Intelligence, Cybersecurity) was addressed across multiple Bloom's levels.
- **Diversity**: The prompts were designed with textual diversity to ensure a variety of responses and to eliminate potential biases within the model's training methods.
- **Grounding**: The prompts were designed to be grounded within specific concepts rather than being broad or non-specific, minimizing hallucinations and maintaining relevance within the generated outputs.
- **Aignment**: Prompts were designed to align the desired outputs with defined learning outcomes within the curriculum. As a result, it significantly increases the opportunity to create applicable learning support and assessment tools for instructional delivery and learning.

Through these design strategies, we were able to create a semantically meaningful and cognitively diverse dataset of questions that aligned with various content areas. The prompts provided a mechanism for linking the unstructured capabilities of the LLM with the structured demands of adaptive student assessments. Thus, the PBAT is a fully developed, educationally supported assessment framework, not merely a commercially developed technical framework.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Pankaj P. Mishra (P.P.M.) conceived the research idea, supervised the overall study, and coordinated the manuscript preparation. V. Venkataramanan (V.V.) contributed to system design, methodology formulation, and technical validation of the proposed approach. Wen Gu (W.G.) provided critical insights, contributed to advanced analysis, and reviewed the manuscript for technical rigor. Keyur Patel (K.P.) and Tirth Patel (T.P.) were involved in data collection, experimentation, and implementation of the proposed models. Asmi Moghe (A.M.) assisted in result analysis, visualization, and documentation. Adwait Patankar (A.P.) supported simulation studies and performance evaluation. P.P.M. drafted the initial version of the manuscript, and all authors reviewed, revised, and approved the final version of the manuscript.

REFERENCES

[1] Y. Chali and S. A. Hasan, "Towards automatic topical question generation," in *Proc. COLING 2012*, 2012, pp. 475–492.

[2] H. Khosravi, S. Sadiq, and D. Gasevic, "Development and adoption of an adaptive learning system: Reflections and lessons learned," in *Proc. 51st ACM Tech. Symp. Comput. Sci. Educ*., 2020, pp. 58–64.

[3] F. Mujtaba and N. R. Mahapatra, "Artificial intelligence in computerized adaptive testing," in *Proc. 2020 Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, IEEE, 2020, pp. 649–654.

[4] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017.

[5] Z. Hou, S. Bi, G. Qi *et al.*, "Syntax-guided question generation using prompt learning," *Neural Comput. Appl*., vol. 36, no. 12, pp. 6271–6282, 2024.

[6] B. Liu, H. Wei, D. Niu *et al.*, "Asking questions the human way: Scalable question-answer generation from text corpus," in *Proc. Web Conf. 2020*, 2020, pp. 202–2043.

[7] N. Scaria, S. Dharani Chenna, and D. Subramani, "Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation," in *Proc. Int. Conf. Artif. Intell. Educ*., Cham: Springer Nature Switzerland, July 2024, pp. 165–179.

[8] J. W. Park, S. J. Park, H. S. Won *et al*., "Large language models are students at various levels: Zero-shot question difficulty estimation," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP 2024*, 2024, pp. 8157–8177.

[9] M. Forehand, "Bloom's taxonomy," *Emerging Perspect. Learn. Teach. Technol.*, vol. 41, no. 4, pp. 47–56, 2010.

[10] J. He-Yueya and A. Singla, "Quizzing policy using reinforcement learning for inferring the student knowledge state," in *Proc. 14th Int. Conf. Educ. Data Mining*, 2021, pp. 533–539.

[11] S. Lamsiyah, A. El Mahdaouy, A. Nourbakhsh *et al.*, "Fine-tuning a large language model with reinforcement learning for educational question generation," in *Proc. Int. Conf. Artif. Intell. Educ.*, Springer, 2024, pp. 424–438.

[12] Z. Jing, Y. Su, and Y. Han, "When large language models meet vector databases: A survey," in *Proc. 2025 Conf. Artif. Intell. x Multimedia (AIxMM)*, IEEE, Feb. 2025, pp. 7–13.

[13] L. Huang *et al*., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, 2025.

[14] J. Dong, "Natural language processing pretraining language model for computer intelligent recognition technology," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 8, pp. 1–12, 2024.

[15] V. Nikolovski, D. Trajanov, and I. Chorbev, "Advancing AI in higher education: A comparative study of large language model-based agents for exam question generation, improvement, and evaluation," *Algorithms*, vol. 18, no. 3, 144, 2025.

[16] M. Heilman and N. A. Smith, "Good question! statistical ranking for question generation," in *Proc. Hum. Lang. Technol.: 2010 Annu. Conf. N. Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 609–617.

[17] Nooralahzadeh, L. Lønning, and L. Øvrelid, "Reinforcement-based denoising of distantly supervised NER with partial annotation," in *Proc. 2nd Workshop Deep Learn. Approaches Low-Resour. NLP (DeepLo 2019)*, 2019, pp. 225–233.

[18] S. Bulathwela, H. Muse, and E. Yilmaz, "Scalable educational question generation with pre-trained language models," in *Proc. Int. Conf. Artif. Intell. Educ.*, July 2023, pp. 327–339.

[19] Q. Zhou, N. Yang, F. Wei *et al.*, "Neural question generation from text: A preliminary study," in *Proc. Nat. CCF Conf. Nat. Lang. Process. Chinese Comput*., Springer, 2017, pp. 662–671.

[20] Z. Li, M. Cukurova, and S. Bulathwela, "A novel approach to scalable and automatic topic-controlled question generation in education," in *Proc. 15th Int. Learn. Analytics Knowl. Conf.*, March 2025, pp. 148–158.

[21] T. Scialom, S. Lamprier, B. Piwowarski *et al.*, "Answers unite! unsupervised metrics for reinforced summarization models," in *Proc. 2019 Conf. Empirical Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3237–3247.

[22] Z. Wang, S. Rao, J. Zhang *et al*., "Diversify question generation with continuous content selectors and question type modeling," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP 2020*, 2020, pp. 2134–2143.

[23] T. Malrick and Y. Corbette, "Multimodal prompt engineering for cross-task vision-language transfer," *J. Comput. Sci. Softw. Appl.*, vol. 5, no. 5, 2025.

[24] Z. Zheng, W. Chao, Z. Qiu, H. Zhu, and H. Xiong, "Harnessing large language models for text-rich sequential recommendation," in *Proc. ACM Web Conf. 2024*, May 2024, pp. 3207–3216.

[25] X. Daomin, "Multi-agent-based e-learning intelligent tutoring system for supporting adaptive learning," in *Proc. 2013 4th Int. Conf. Intell. Syst. Des. Eng. Appl.*, IEEE, 2013, pp. 393–397.

[26] D. Lohr, M. Berges, A. Chugh, M. Kohlhase, and D. Müller, "Leveraging large language models to generate course-specific semantically annotated learning objects," *J. Comput. Assisted Learn*., vol. 41, no. 1, e13101, 2025.

[27] M. Aigerim, T. Arailym, N. Aliya, S. Adai, and S. E. Seker, "A systematic evaluation of large language models and retrieval-

augmented generation for the task of kazakh question answering," *Information*, vol. 16, no. 11, p. 943, 2025.
[28] M. J. Zyphur and F. L. Oswald, "Bayesian estimation and inference: A user's guide," *J. Manage.*, vol. 41, no. 2, pp. 390–420, 2015.