Volume 11, Issue 8, August – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

Artificial Intelligence-Driven Adaptive Assessment Systems: A Systematic Survey

Srusti S. Deshmukh[1]; Dr. B. R. Mohan[2]

[1]Department of Computer Science and Engineering, East West Institute of Technology
[2]Professor, Department of Computer Science and Engineering, East West Institute of Technology
Bengaluru, India

Publication Date: 2026/08/24

Abstract: Modern educational assessment faces critical challenges in capturing fine-grained student proficiency through static, fixed-form testing methods that impose uniform burdens and overlook dynamic learning trajectories. To address these limitations, artificial intelligence has emerged as a transformative paradigm, shifting traditional static testing toward intelligent, real-time adaptive ecosystems capable of continuous student modeling and personalized evaluation. This paper presents a systematic survey of artificial intelligence-driven adaptive assessment systems, synthesized through a rigorous PRISMA-based review methodology. The study systematically examines both foundational psychometric milestones and recent technological advances across peer-reviewed literature. A novel end-to-end taxonomy is introduced, categorizing the literature into five core pillars: measurement models spanning classical test theory to neural cognitive diagnosis, adaptive item selection algorithms leveraging psychometric information and reinforcement learning, automated item generation driven by transformer models and large language models, automated scoring mechanisms utilizing contextual embeddings, and underlying system architectures paired with learning analytics. Furthermore, this survey provides a comparative analysis of existing methods, identifies critical open research challenges regarding algorithmic bias and model interpretability, and outlines a comprehensive future research roadmap. Ultimately, this work highlights the critical significance of developing trustworthy, explainable, and intelligent adaptive assessment systems for modern education.

Keywords: *Artificial Intelligence, Adaptive Assessment, Computerized Adaptive Testing, Item Response Theory, Knowledge Tracing, Neural Cognitive Diagnosis, Educational Data Mining, Large Language Models.*

How to Cite: Srusti S. Deshmukh; Dr. B. R. Mohan (2026) Artificial Intelligence-Driven Adaptive Assessment Systems: A Systematic Survey. *International Journal of Innovative Science and Research Technology*, 11(8), 1500-1519. https://doi.org/10.38124/ijisrt/26aug610

I. INTRODUCTION

➢ *Background of Educational Assessment*

Systematic educational evaluation serves as a cornerstone of instructional design, providing the primary empirical foundation for tracking cognitive development and certifying academic achievement [1, 2]. Historically, measurement practices have evolved from rudimentary score-reporting tools into structured diagnostic frameworks capable of mapping learner proficiency and guiding pedagogical decision-making [1]. By systematically capturing student performance, these instruments bridge the critical gap between curriculum delivery and learning outcomes [1, 2]. This evaluative process allows educators to monitor progress, adapt instructional strategies, and ensure that educational standards are maintained across diverse cohorts [2]. However, fulfilling these diagnostic objectives depends heavily on the structural capability of the testing instruments themselves [1]. When legacy evaluation tools are deployed at scale, their structural constraints quickly become apparent, setting the stage for an examination of the systemic shortcomings inherent in conventional, fixed-form measurement [1, 2].

➢ *Limitations of Traditional Assessment*

Standard educational evaluation has long depended on fixed-form assessments, in which every examinee completes a predetermined, static sequence of test items regardless of individual proficiency [1, 2]. This uniform delivery model introduces profound structural and operational constraints [1]. Because item difficulty parameters remain invariant during testing, high-ability students frequently encounter ceiling effects that cap observable performance, whereas struggling learners experience floor effects and disproportionate testing fatigue [1, 3]. Consequently, traditional instruments exhibit measurement inefficiency, often requiring extended test lengths to secure reliable proficiency estimates across heterogeneous populations [2]. Furthermore, these legacy formats fail to provide granular diagnostic feedback regarding specific cognitive gaps or longitudinal learning trajectories [1]. Such vulnerabilities become acutely problematic within modern digital educational settings where student capabilities vary widely [2]. The inherent inability of static testing paradigms to dynamically adjust item difficulty highlights the fundamental limits of fixed-form evaluation [1, 2]. Recognizing these critical performance bottlenecks naturally

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

motivates the historical shift toward adaptive assessment methodologies designed to tailor test delivery to individual proficiency [2].

➢ *Evolution to Adaptive Assessment*

To overcome the systemic inefficiencies and rigid constraints of static fixed-form evaluation, educational measurement evolved toward personalized paradigms known as Computerized Adaptive Testing (CAT) [2]. Serving as a critical evolutionary bridge, CAT shifts evaluation from uniform item delivery to dynamic, proficiency-tailored testing routines that continuously adjust to individual learner capability [2]. This developmental trajectory from classical psychometric methods to intelligent assessment ecosystems is conceptualized in Figure 1. Evolution from Traditional Assessment to Artificial Intelligence-Driven Adaptive Assessment Systems.

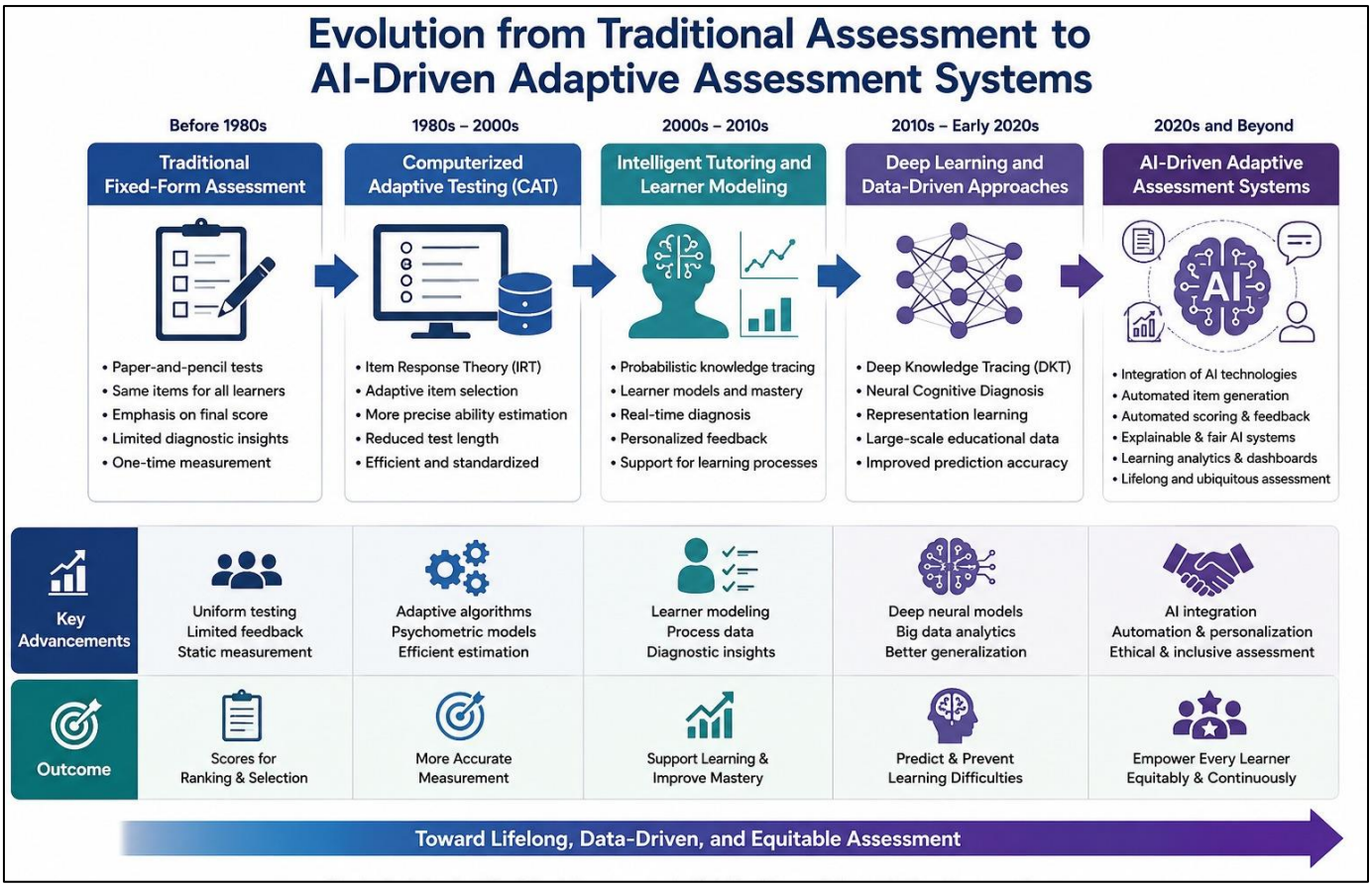

Fig 1 Evolution from Traditional Assessment to Artificial Intelligence-Driven Adaptive Assessment Systems

The operational foundation of CAT relies fundamentally upon Item Response Theory (IRT), which establishes invariant item parameters and models the nonlinear probabilistic relationship between latent student traits and correct item endorsement [1, 2]. By continuously updating ability estimates and utilizing Fisher information functions, CAT selects subsequent test items that maximize measurement precision at the examinee's exact proficiency level, thereby substantially reducing the overall testing burden [2]. Despite these efficiency gains, traditional CAT frameworks remain constrained by strict psychometric assumptions—such as unidimensionality and local independence—which restrict their adaptability when confronting multidimensional cognitive structures and complex, evolving learner behaviors [3]. These structural boundaries of classical psychometric adaptation highlight the necessity for advanced computational methods, naturally motivating the integration of artificial intelligence into contemporary adaptive assessment systems.

➢ *Emergence of Artificial Intelligence in Adaptive Assessment*

While Computerized Adaptive Testing advanced psychometric efficiency, traditional algorithms face notable structural constraints when deployed in modern digital learning environments characterized by massive, multi-modal student interaction streams. The exponential expansion of online educational platforms has generated unprecedented volumes of behavioral data, exposing the limitations of static psychometric models in capturing continuous cognitive changes over extended learning trajectories. Artificial intelligence has increasingly extended the evolution of adaptive assessment by complementing rather than replacing classical psychometric foundations, while integrating statistical measurement theory with advanced representation learning. Rather than operating as isolated techniques, modern computational methods synthesize an interconnected ecosystem encompassing deep knowledge tracing for continuous student modeling, neural cognitive diagnosis for attribute mastery, reinforcement learning for intelligent item

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

selection, transformer models for automated item generation, and contextual embeddings for automated scoring. This fully integrated pipeline transforms adaptive assessment into a cohesive operational framework. Nevertheless, current research often studies these complementary capabilities—including learner modeling, item selection, automated content generation, automated scoring, and educational analytics—independently. By unifying these diverse research directions through a coherent end-to-end taxonomy, this survey addresses this fragmentation and establishes a comprehensive foundation for future development.

➢ *Research Gap*

While significant progress has been achieved in adaptive assessment research, existing literature exhibits notable fragmentation. Previous surveys have provided valuable insights into isolated components, such as Computerized Adaptive Testing, Item Response Theory, knowledge tracing models, cognitive diagnosis, automated item generation, large language models, and automated scoring [1, 2]. However, these topics are predominantly investigated as independent domains rather than as parts of a cohesive, integrated adaptive assessment ecosystem. Consequently, there remains a critical absence of a comprehensive survey that systematically connects learner modeling, adaptive decision-making, automated content generation, response evaluation, and underlying system architecture within one unified framework. Furthermore, prior reviews offer limited critical discussion concerning cross-cutting operational challenges, including model explainability, algorithmic fairness, cross-domain interoperability, and real-world deployment constraints. This survey addresses these critical gaps by synthesizing contemporary research through a rigorous PRISMA-based methodology, establishing a unified end-to-end taxonomy, providing deep comparative analyses, and outlining a structured research roadmap, as detailed in subsection I.F. Contributions of This Survey.

➢ *Contributions of This Survey*

To address the aforementioned research gaps, this paper provides a comprehensive evaluation of contemporary advancements in educational technology. Specifically, the primary contributions of this systematic survey are summarized as follows:

- A comprehensive systematic survey of Artificial Intelligence-driven adaptive assessment systems, synthesized through a rigorous PRISMA-based review methodology.
- A unified end-to-end taxonomy that systematically integrates five core pillars: measurement models, adaptive item selection, automated item generation, automated scoring, and system architecture paired with learning analytics.
- A critical comparative analysis of existing approaches, evaluating their underlying theories, mathematical foundations, interpretability, computational complexity, scalability, and educational applicability.
- A thorough synthesis of emerging critical challenges, encompassing algorithmic fairness, model explainability, cross-domain interoperability, cold-start problems, real-world deployment constraints, and ethical considerations.
- A structured future research roadmap that highlights promising methodological directions for next-generation intelligent adaptive assessment systems.

Collectively, these contributions establish a cohesive framework for understanding the current landscape and future trajectory of intelligent evaluation, setting the foundation for the systematic investigation detailed next in Section II. Survey Methodology.

II. SURVEY METHODOLOGY

➢ *PRISMA Framework*

Establishing a rigorous, transparent, and reproducible methodology is essential for conducting a high-quality systematic survey of artificial intelligence-driven educational assessment in complex modern environments. To achieve this methodological standard, this study adopts the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines [4]. The literature selection process is systematically structured across four sequential phases: identification, screening, eligibility, and inclusion. Implementing this rigorous protocol effectively minimizes selection bias, ensures objective evaluation, and enhances the overall transparency and reproducibility of the survey execution across diverse databases. The complete literature selection process is illustrated in Fig. 2. This robust methodological framework provides the essential foundation for formulating the core research questions that define the scope of the survey, leading directly into subsection II.B. Research Questions.

➢ *Research Questions*

Consistent with the PRISMA 2020 methodology, clearly defined research questions establish the scope, direction, and analytical focus of this systematic survey [4]. To examine the multidimensional landscape of artificial intelligence-driven educational evaluation, this study addresses five core research questions:

- RQ1: What are the principal measurement models employed in artificial intelligence-driven adaptive assessment systems?
- RQ2: How do modern adaptive item selection techniques improve personalized assessment compared with conventional psychometric approaches?
- RQ3: How are artificial intelligence techniques utilized for automated item generation and automated scoring within adaptive assessment systems?
- RQ4: What computational architectures, learning analytics frameworks, and intelligent educational infrastructures support end-to-end adaptive assessment?
- RQ5: What are the current research challenges, limitations, and promising future directions for artificial intelligence-driven adaptive assessment systems?

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

These research questions collectively guided the literature search, study screening, taxonomy development, comparative analysis, synthesis of findings, and identification of future research directions. Together, they directly informed the design and implementation of the literature search strategy.

➢ *Search Strategy*

To identify relevant literature addressing the formulated research questions, a comprehensive electronic search strategy was executed across major scholarly databases, including IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Scopus. The search string design incorporated precise keyword combinations and Boolean operators (such as AND, OR) to capture the convergence of artificial intelligence and adaptive educational assessment. The primary systematic search targeted peer-reviewed journal articles and conference proceedings published between 2015 and 2026 to capture recent technological advancements in AI-driven adaptive assessment. However, seminal publications predating 2015 were intentionally included when they established the fundamental theoretical and psychometric foundations required to understand modern systems, including classical test theory, Item Response Theory, Computerized Adaptive Testing, and Bayesian Knowledge Tracing. Database-specific filters were applied where appropriate to ensure that only high-quality scholarly publications reporting empirical implementations, psychometric models, and computational algorithms were retrieved. This structured querying protocol established the robust baseline data pool required for subsequent eligibility screening.

➢ *Inclusion Criteria*

Following the identification phase, retrieved records underwent a rigorous evaluation to determine study eligibility based on explicit inclusion criteria. To ensure high academic quality and relevance, studies were required to meet the following conditions: first, publication in peer-reviewed journals or reputable international conference proceedings; second, explicit focus on artificial intelligence-driven methodologies applied to adaptive educational assessment, computerized adaptive testing, knowledge tracing, or automated item generation and scoring; third, clear methodological rigor, including well-defined algorithmic architectures, mathematical formalizations, or robust psychometric frameworks; fourth, provision of empirical validation, simulation-based evaluation, or comparative performance analysis using benchmark educational datasets; and fifth, publication in English to maintain analytical consistency. These objective parameters ensured that only high-integrity, directly applicable scholarship advanced to the full-text review stage.

➢ *Exclusion Criteria*

To maintain focus and prevent methodological scope creep, clear exclusion criteria were enforced during the screening and eligibility phases. Studies were excluded if they represented duplicate records, non-peer-reviewed materials, editorials, commentaries, abstracts-only publications, textbooks, doctoral dissertations, or gray literature. Furthermore, publications failing to provide full-text accessibility or written in languages other than English were omitted. Research focusing exclusively on traditional, static fixed-form testing without computational adaptation, or general machine learning applications lacking direct pedagogical and assessment contexts, was also systematically filtered out. Enforcing these transparent exclusion parameters ensured analytical consistency, reduced selection bias, and restricted the final review corpus exclusively to relevant peer-reviewed contributions to intelligent adaptive assessment systems.

➢ *Literature Selection Process*

The retrieved records underwent a structured multi-stage screening process following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) methodology [4]. The review pipeline followed a systematic sequence: initial identification of records across electronic databases, systematic removal of duplicate entries, screening of titles and abstracts against inclusion and exclusion criteria, and rigorous full-text eligibility assessment, culminating in the final inclusion of relevant studies. Specifically, 6,475 records were initially retrieved, followed by duplicate removal and sequential screening based on the predefined eligibility criteria. A total of 333 studies were included in the qualitative synthesis, while 27 additional studies were identified through backward and forward snowballing, resulting in 360 studies included in the systematic review. The complete literature selection workflow is illustrated in Fig. 2. This systematic filtering process ensured methodological transparency, reproducibility, and consistency while minimizing selection bias. The resulting collection of eligible studies constitutes the analytical foundation for the taxonomy, comparative analyses, and technical discussions presented throughout the remainder of the survey, leading directly into Section III: A Taxonomy of Artificial Intelligence-Driven Adaptive Assessment Systems.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

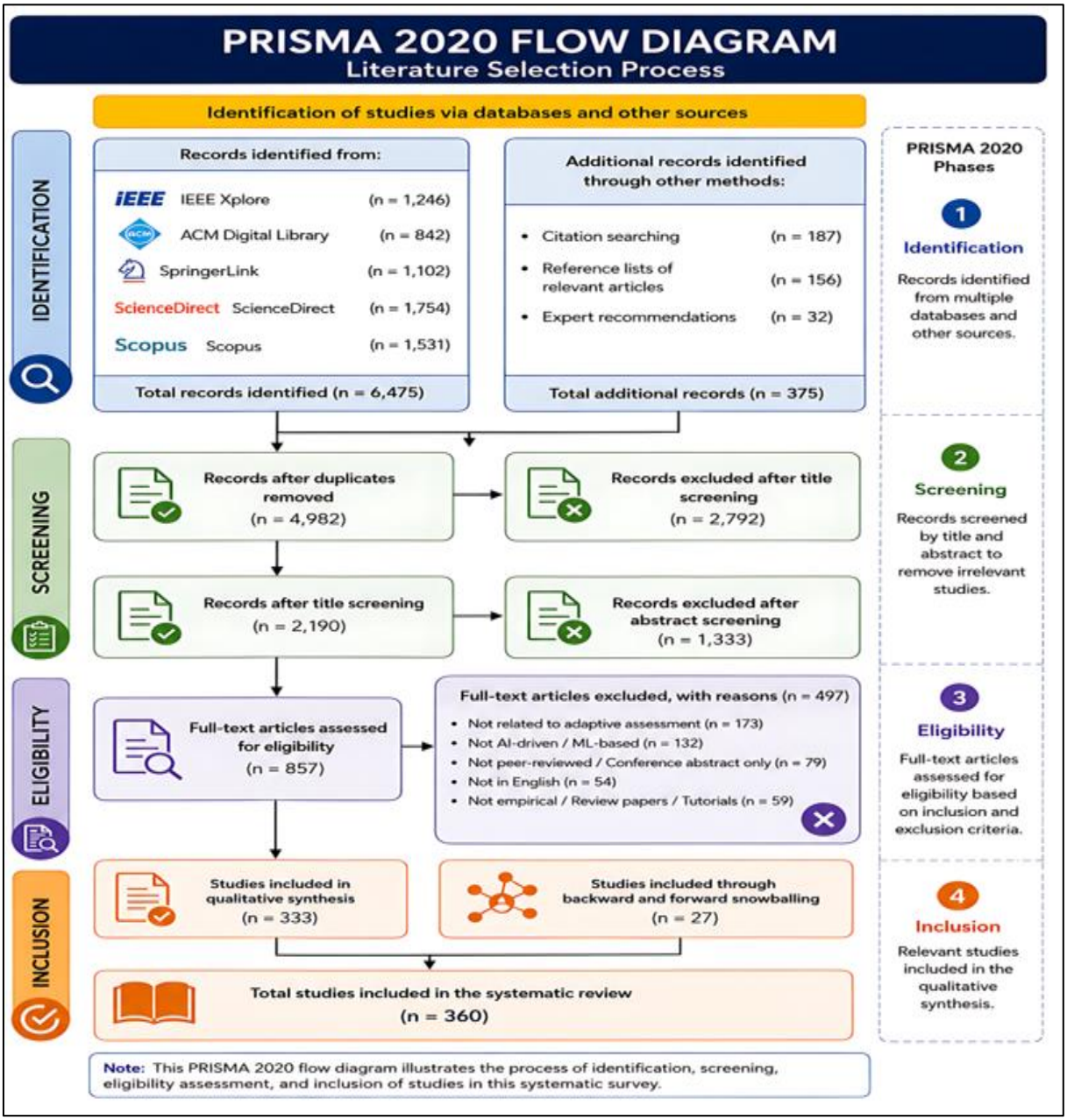

Fig 2 PRISMA 2020 Flow Diagram of the Literature Selection Process

III. A TAXONOMY OF ARTIFICIAL INTELLIGENCE- DRIVEN ADAPTIVE ASSESSMENT SYSTEMS

➢ *Proposed End-to-End Taxonomy*

Existing research concerning intelligent educational assessment often investigates individual components in isolation, resulting in a fragmented scholarly landscape. To overcome this systemic limitation, a comprehensive and unified framework is essential to connect disparate technical developments into a cohesive operational workflow. This survey introduces a unified end-to-end taxonomy that systematically organizes the complete lifecycle of artificial intelligence-driven adaptive assessment systems. The proposed taxonomy encompasses five core pillars: measurement models, adaptive item selection, automated item generation, automated scoring, and system architecture paired with learning analytics. The proposed taxonomy is illustrated in Fig. 3. By structuring the literature through this end-to-end lens, the taxonomy integrates learner profiling, content delivery, response evaluation, and infrastructural management into one unified operational lifecycle. This holistic perspective provides a structured foundation for analyzing modern educational technologies, leading naturally into the examination of measurement.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

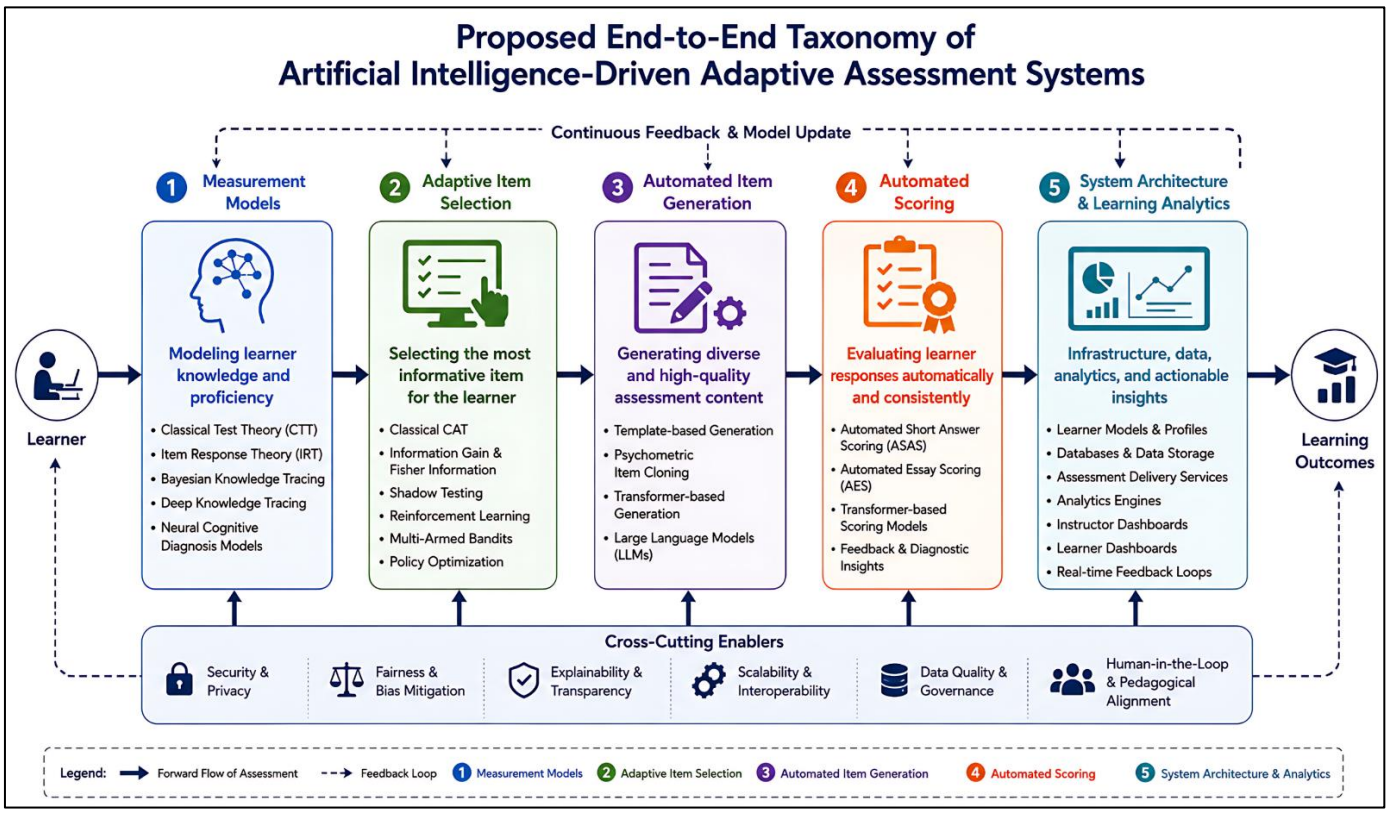

Fig 3 Proposed End-to-End Taxonomy of Artificial Intelligence-Driven Adaptive Assessment Systems

➢ *Measurement Models*

Measurement models serve as the foundational computational core of intelligent adaptive assessment by mathematically operationalizing learner proficiency and tracking cognitive states over time [1, 2, 5]. Within this pillar, methodologies span classical psychometric foundations to advanced representation learning paradigms. Classical Test Theory and Item Response Theory establish foundational probabilistic relationships between latent examinee traits and item responses [1, 2]. To capture dynamic longitudinal changes, Bayesian Knowledge Tracing models mastery states as discrete Markov chains [5], whereas deep knowledge tracing leverages recurrent neural architectures to model student learning trajectories from sparse interaction logs [6]. Furthermore, neural cognitive diagnosis frameworks integrate deep learning with psychometric principles to deliver fine-grained, multidimensional attribute mastery profiles [5, 6]. Rather than operating independently, these diverse models quantify student knowledge states and track cognitive evolution, providing the essential precision required to drive adaptive item selection.

➢ *Adaptive Item Selection*

Building upon the learner representations established by measurement models, adaptive item selection algorithms determine the optimal next assessment item for each individual examinee. Traditional Computerized Adaptive Testing relies primarily on information-theoretic principles, utilizing Fisher information functions to select items that maximize measurement precision at the current proficiency estimate [2]. Modern intelligent frameworks extend these classical paradigms by incorporating reinforcement learning and sequential decision-making policies, enabling systems to optimize long-term educational objectives, balance exploration and exploitation, and adapt to evolving learner behaviors [16, 17]. By dynamically updating the estimated student state following each interaction, these advanced item selection mechanisms substantially reduce testing length while maximizing diagnostic efficiency and personalization. This seamless integration of psychometric rigor with sequential optimization ensures precise, individualized evaluation, determining what should be assessed next and thereby motivating the need for scalable automated content generation.

➢ *Automated Item Generation*

Maintaining an efficient adaptive testing environment requires a continuous, scalable supply of high-quality test items, which often strains traditional manual authoring pipelines. To overcome this critical bottleneck, automated item generation has emerged as an essential component, shifting content creation from static item banks to dynamic, algorithmic generation [18]. While early implementations relied primarily on template-based models and psychometric item-cloning rules, recent advancements transition heavily toward transformer-based neural generation methods. Large language models now serve as modern solutions for automated question Generation, distractor creation, and contextual assessment development [19]. These generative frameworks offer substantial advantages in scalability, diversity, and targeted personalization, ensuring that item pools remain secure and adaptable to varied learner cohorts. By delivering rich content efficiently, automated item generation prepares the ecosystem for subsequent response evaluation, transitioning naturally toward automated scoring.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

➢ *Automated Scoring*

Once examinees complete the delivered assessment items, efficient and accurate evaluation is required to close the learning loop. Automated scoring leverages advanced computational techniques to evaluate learner responses across diverse formats [20]. Initial developments focused on Automated Short Answer Scoring to evaluate concise text responses, which subsequently expanded toward sophisticated Automated Essay Scoring systems capable of analyzing complex compositions. Modern systems utilize transformer-based contextual embeddings to capture semantic meaning, syntax, and discourse properties beyond simple lexical matching [21]. These scoring frameworks deliver critical advantages, including high grading consistency, rapid feedback delivery, operational scalability, and personalized diagnostic insights. By processing complex learner submissions efficiently without manual latency, automated scoring connects evaluation directly to the broader backend infrastructure, transitioning smoothly toward system architecture and learning analytics.

➢ *System Architecture and Learning Analytics*

The operational execution of intelligent adaptive assessment depends critically upon robust backend architectures and data-driven learning analytics frameworks [22]. Modern computational infrastructures integrate distributed backend services, secure student databases, high-performance assessment delivery services, and real-time analytics engines to handle continuous interaction streams [23]. These components support dedicated instructor dashboards and learner dashboards that visualize performance trends, alongside feedback loops that update learner models dynamically [24]. Learning analytics transforms raw interaction logs into actionable educational insights by monitoring engagement patterns and system efficacy. By maintaining a scalable architectural perspective, these platforms ensure seamless coordination across all assessment phases, paving the way for a unified evaluation of the complete framework.

➢ *Discussion of the Proposed Taxonomy*

The proposed taxonomy synthesizes the five core pillars—measurement models, adaptive item selection, automated item generation, automated scoring, and system architecture—into a coherent operational lifecycle [11]. Previous literature frequently investigates these components in isolation, treating learner profiling, content generation, and evaluation as disconnected domains. By integrating these areas, the taxonomy demonstrates how learner modeling, adaptive decision-making, content generation, automated evaluation, and system infrastructure function as a continuous ecosystem. This structured organization provides a unified analytical perspective rather than introducing isolated algorithms, supporting systematic comparison across subsequent sections. Establishing this integrated framework ensures a clear structural roadmap, leading naturally into the detailed technical analysis of measurement models presented in Section IV.

IV. MEASUREMENT MODELS

Measurement models form the quantitative backbone of intelligent adaptive assessment systems, providing the mathematical mechanisms required to estimate examinee proficiency and track cognitive development over time [1], [2]. These frameworks establish the relationship between observable learner responses and unobservable latent traits, thereby ensuring that adaptive item selection, personalized feedback, and diagnostic inference are grounded in rigorous psychometric theory. As educational assessment has evolved, measurement models have progressed from classical statistical paradigms to probabilistic graphical models and modern deep learning architectures. Table I summarizes the principal characteristics, strengths, limitations, and computational properties of the major educational measurement models discussed in this section.

➢ *Classical Test Theory*

Classical Test Theory (CTT) represents the historical foundation of educational measurement and provides one of the earliest quantitative frameworks for evaluating examinee performance [1], [2]. CTT assumes that an individual's observed score consists of two components: the true score reflecting actual ability and a random measurement error arising from uncontrollable external factors. This relationship is expressed as:

$$X = T + E \quad (1)$$

Where X denotes the observed test score, T represents the examinee's true score, and E corresponds to random measurement error.

CTT assumes that measurement errors have an expected value of zero, are randomly distributed, and remain uncorrelated with true scores. Owing to its conceptual simplicity and modest computational requirements, CTT has been widely adopted in standardized educational testing and classroom assessment. Nevertheless, the framework suffers from important limitations. Item statistics such as difficulty and discrimination are sample-dependent, whereas examinee scores are test-dependent, preventing reliable comparisons across different examinations or populations. These structural shortcomings motivated the development of Item Response Theory, which provides parameter invariance and a more robust probabilistic framework for educational measurement.

➢ *Item Response Theory*

Item Response Theory (IRT) addresses the limitations of Classical Test Theory by modeling the probabilistic relationship between an examinee's latent ability and the characteristics of individual assessment items [1], [2]. Unlike CTT, IRT estimates both examinee ability and item parameters on a common latent scale, thereby enabling comparisons across different test forms.

Depending on the number of item parameters considered, IRT is commonly categorized into the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models. Among these, the 3PL

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

model is the most comprehensive because it simultaneously accounts for item discrimination, difficulty, and pseudo-guessing.

$$P_{i(\theta)} = c_i + \frac{1-c_i}{1+e^{-a_i(\theta-b_i)}} \quad (2)$$

Where

- Denotes the probability that an examinee with latent ability answers item correctly.
- Is the item discrimination parameter.
- Is the item difficulty parameter.
- Represents the pseudo-guessing parameter.
- Denotes the latent proficiency level of the examinee.

In the 1PL model, discrimination is assumed identical across all items and the pseudo-guessing parameter is fixed to zero. The 2PL model estimates discrimination and difficulty while maintaining. The 3PL model additionally estimates guessing behavior, making it particularly suitable for multiple-choice assessments.

IRT further employs Fisher Information to quantify the precision with which individual items estimate learner proficiency, enabling Computerized Adaptive Testing systems to dynamically select the most informative subsequent question. Despite these advantages, traditional IRT generally assumes unidimensional latent ability and static knowledge states, limiting its ability to capture continuous learning over time. These limitations motivated the emergence of Bayesian Knowledge Tracing.

➢ *Bayesian Knowledge Tracing*

Although Item Response Theory accurately estimates learner proficiency at a specific point in time, intelligent tutoring environments require continuous modeling of evolving knowledge states. Bayesian Knowledge Tracing (BKT) addresses this challenge by representing student learning as a Hidden Markov Model in which each knowledge component exists in either a mastered or unmastered state [5].

BKT is governed by four probabilistic parameters:

- $P(L_0)$: initial probability that a learner has already mastered the skill,
- $P(T)$: probability of learning after an interaction,
- $P(G)$: probability of correctly answering despite lacking mastery (guessing),
- $P(S)$: probability of incorrectly answering despite mastery (slipping).

Following each learner response, the probability of mastery is updated using Bayes' theorem.

$$P(L_t \mid Correct_t) = \frac{P(L_{\{t-1\}})(1-P(S))}{P(L_{\{t-1\}})(1-P(S))+(1-P(L_{\{t-1\}}))P(G)\}} \quad (3)$$

Where denotes the mastery probability before observing the current response. After Bayesian updating, the transition parameter accounts for possible learning that occurs during the interaction.

BKT offers strong probabilistic interpretability and computational efficiency, making it one of the most widely deployed learner modeling techniques in intelligent tutoring systems. However, the framework assumes homogeneous learning behavior across students and models each skill independently, preventing it from capturing complex dependencies among multiple knowledge concepts. These limitations inspired the development of Deep Knowledge Tracing.

➢ *Deep Knowledge Tracing*

Deep Knowledge Tracing (DKT) replaces the manually designed probabilistic assumptions of Bayesian Knowledge Tracing with data-driven neural sequence models capable of learning complex temporal patterns directly from educational interaction data [6]. Rather than explicitly modeling learner transitions, DKT employs recurrent neural networks to learn latent representations of student knowledge from sequences of historical responses.

The hidden state evolution of the learner model is expressed as:

$$h_t = \{LSTM\}(x_t, h_{\{t-1\}}) \quad (4)$$

Where

- Represents the input embedding corresponding to the exercise-response pair at time step t,
- Denotes the hidden representation from the previous interaction,
- Represents the Long Short-Term Memory transition function.

Recent extensions include Deep Knowledge Value Memory Networks (DKVMN), attention-based knowledge tracing, and transformer-based knowledge tracing architectures, all of which improve long-range dependency modeling through explicit memory mechanisms and self-attention [6].

Although DKT substantially improves prediction accuracy and eliminates the need for manual knowledge component engineering, its highly nonlinear architecture reduces interpretability, motivating research into neural cognitive diagnosis models that combine deep learning with explicit psychometric reasoning.

➢ *Neural Cognitive Diagnosis Models*

Neural Cognitive Diagnosis (Neural CD) integrates deep representation learning with cognitive diagnosis theory to estimate learner mastery across multiple knowledge attributes simultaneously. Representative models include NeuralCD, Deep-IRT, Reciprocal Cognitive Diagnosis (RCD), and Monotonic Neural Cognitive Diagnosis.

The learner-item interaction is generally formulated as:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

$$y_{\{ij\}} = f(\theta_i, \beta_j, Q) \quad (5)$$

Where

- Denotes the predicted interaction outcome,
- Represents the multidimensional latent proficiency vector of learner i,
- Denotes the parameter vector describing assessment item j,
- Represents the Q-matrix encoding prerequisite relationships between assessment items and knowledge concepts.

Unlike traditional psychometric models that estimate a single proficiency score, neural cognitive diagnosis provides fine-grained mastery estimates for individual concepts while preserving psychometric interpretability. These models effectively combine the expressive learning capacity of deep neural networks with the diagnostic transparency required for educational decision-making, thereby establishing a bridge between classical measurement theory and modern artificial intelligence.

➢ *Comparative Discussion*

The evolution of educational measurement models reflects a continuous progression from simple statistical score estimation toward sophisticated, multidimensional learner modeling. Table 1 summarizes the comparative characteristics of the principal measurement models discussed in this section.

Table 1 Comparative Characteristics of Principal Measurement Models

Model	Computational Basis	Dynamics & Granularity	Interpretability
CTT	True-score additive model ($X = T + E$)	Static; test-level score	High
IRT	Probabilistic logistic item response curves (1PL,2PL, 3PL)	Static; latent ability (θ)	High
BKT	Hidden Markov Chains with Bayesian updating	Dynamic; binary skill mastery	Moderate
DKT	Recurrent neural networks (LSTM/GRU)	Dynamic; continuous sequence tracking	Low (Black-box)
Neural CD	Neural-psychometric Q-matrix integration	Dynamic; multidimensional mastery	Moderate-High

Classical Test Theory establishes the foundational true-score paradigm but suffers from sample-dependent parameter estimation and limited diagnostic capability [1], [2]. Item Response Theory overcomes these limitations through latent trait modeling and parameter invariance, providing the mathematical foundation for Computerized Adaptive Testing [2]. Bayesian Knowledge Tracing extends measurement into temporal learning environments using probabilistic state transitions but assumes independent skills and homogeneous learning behaviors [5]. Deep Knowledge Tracing further advances learner modeling by capturing complex sequential dependencies through recurrent neural networks, although this improvement comes at the expense of interpretability [6]. Finally, Neural Cognitive Diagnosis integrates psychometric theory with deep learning to provide fine-grained, multidimensional learner diagnosis while maintaining educational interpretability.

Beyond improving predictive performance, this progression reflects a gradual transition from aggregate score estimation toward interpretable learner representations capable of supporting personalized educational interventions. Collectively, these increasingly sophisticated learner models provide the diagnostic foundation for adaptive item selection algorithms, which are examined in the following section.

V. ADAPTIVE ITEM SELECTION

➢ *Classical Computerized Adaptive Testing*

Classical Computerized Adaptive Testing (CAT) transitions educational assessment from static, fixed-form evaluations to dynamic, individualized testing paradigms [2]. Unlike traditional tests where every examinee receives a predetermined sequence of questions, CAT tailors the examination path to the individual's estimated proficiency in real time. The operational workflow begins with the construction of a calibrated item bank containing items whose psychometric parameters are pre-estimated using Item Response Theory [1, 2]. Initialization typically assigns an initial ability estimate θ to the examinee, often using a neutral baseline or available prior assessment information to mitigate cold-start effects. An initial item of intermediate difficulty is selected from the bank and administered. Upon response collection, the examinee's proficiency is re-estimated using maximum likelihood estimation or Bayesian methods [2].

This iterative cycle—selecting the next optimal item based on the updated ability estimate, administering the item, and re-estimating proficiency—continues dynamically. The testing process terminates when a predefined stopping criterion is met, such as reaching a target standard error of measurement or a fixed test length, after which a final proficiency estimate is generated [2]. CAT offers substantial advantages over fixed-form assessment, including significantly reduced test lengths, improved measurement efficiency, individualized item difficulty, and high estimation precision across diverse populations [1, 2]. However, classical CAT encounters notable operational limitations, including heavy dependence on large, rigorously calibrated item banks, challenges in item exposure control and content balancing, cold-start initialization hurdles, and strict reliance on underlying psychometric assumptions like unidimensionality [2, 3]. Despite these constraints, classical CAT establishes the essential methodological foundation for intelligent adaptive item selection.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

➢ *Fisher Information-Based Item Selection*

The mathematical core of classical adaptive item selection relies heavily on information theory to maximize measurement precision at the examinee's current proficiency level [2]. Fisher Information serves as the primary quantitative criterion for determining the suitability of candidate items. For an item i, the general Fisher Information expression for a Bernoulli item-response model evaluated at a given ability estimate θ is expressed as

$$I_i(\theta) = \frac{\left(P'_i(\theta)\right)^2}{P_i(\theta)\left(1 - P_i(\theta)\right)} \quad (6)$$

Where $P_i(\theta)$ denotes the probability of a correct response and $P'_i(\theta)$ represents its first derivative with respect to θ [1, 2]. The specific item information function is derived according to the selected IRT model and its item parameters, such as those formulated in the 2PL and 3PL frameworks introduced in Section IV. Here, θ represents the current ability estimate, while item discrimination and item difficulty parameters dictate the shape and peak of the information curve. CAT algorithms conventionally select the unadministered item that maximizes $I_i(\theta)$ at the examinee's current proficiency estimate [2]. This direct relationship indicates that higher Fisher Information corresponds directly to lower estimation uncertainty and higher measurement precision. Nevertheless, purely myopic information-maximization strategies present distinct limitations. Because they focus exclusively on local optimization for the current item, they frequently cause severe item exposure imbalances, violate content coverage constraints, compromise test security through the repeated selection of informative items, and fail to optimize long-term learning or diagnostic objectives [2, 3]. These operational constraints highlight the necessity for sequential decision-making frameworks, naturally motivating the transition toward reinforcement learning approaches.

➢ *Reinforcement Learning for Adaptive Item Selection*

To overcome the myopic limitations of classical information-maximization strategies, modern adaptive assessment conceptualizes item selection as a sequential decision-making problem governed by reinforcement learning [16, 17]. Rather than optimizing only for immediate measurement precision, reinforcement learning frameworks model the interaction between an adaptive system and a learner as a Markov decision process. In this formulation, the environment state represents the evolving learner knowledge or proficiency representation, the action corresponds to the selection of the next assessment item from the available bank, and the reward reflects a composite metric encompassing measurement precision, learning gain, diagnostic value, or long-term educational benefit [16]. The system learns an optimal policy mapping examinee states to item-selection actions, carefully balancing the exploration of unfamiliar concept spaces with the exploitation of known proficiency boundaries [16, 17]. Conceptually, representative computational architectures span tabular Q-learning, Deep Q-Networks for high-dimensional state representations, policy-gradient methods, and actor-critic frameworks [16, 17]. These advanced models deliver significant advantages, including long-horizon optimization, highly personalized assessment trajectories, dynamic adaptation to changing learner states, and the seamless integration of multi-objective pedagogical criteria [16, 17]. However, reinforcement learning methods introduce distinct challenges, including complex reward function design, training instability, high data requirements, computational complexity, and ensuring educational safety and validity during exploratory phases [16]. These challenges underscore the value of alternative sequential paradigms such as multi-armed bandits.

➢ *Multi-Armed Bandit and Sequential Decision Approaches*

Multi-armed bandit formulations provide a streamlined methodological compromise for adaptive item selection under uncertainty, bridging classical psychometric testing and full reinforcement learning frameworks [16, 17]. In a multi-armed bandit model, each available test item or content module is treated as an independent action or arm. The system must continuously balance exploration—testing items with uncertain efficacy or parameter estimates—with exploitation, selecting items known to align closely with the examinee's estimated ability level [16, 17]. Contextual bandit extensions further refine this mechanism by incorporating rich contextual information, including real-time learner proficiency, historical response patterns, item difficulty parameters, current cognitive states, and prior item exposure frequencies [16].

Methodologically, bandit algorithms differ from classical CAT by handling uncertainty through explicit exploration-exploitation trade-offs rather than strictly maximizing local Fisher Information, and they differ from full reinforcement learning by focusing on immediate or near-term reward optimization rather than full long-horizon Markov decision trajectories [16, 17]. While bandit approaches offer robust adaptability, operational simplicity, and efficient handling of cold-start scenarios, they also face limitations such as reward function misspecification, handling non-stationary learner behavior, maintaining item exposure equity, ensuring demographic fairness, and managing exploration costs in live educational environments [16].

➢ *Comparative Discussion*

Table 2 summarizes the principal characteristics of adaptive item selection approaches. A critical synthesis of the methodologies reviewed in Section V reveals a distinct evolutionary trajectory spanning psychometric optimization, sequential optimization, and long-horizon intelligent decision-making. Classical CAT and Fisher Information-based item selection prioritize immediate local measurement precision and parameter invariance, offering exceptional psychometric rigor and interpretability at the cost of long-term optimization and flexibility [1, 2]. Conversely, multi-armed bandits introduce principled exploration-exploitation under uncertainty, accommodating contextual student features while maintaining computational tractability [16, 17]. Full reinforcement learning models extend these capabilities toward comprehensive long-horizon optimization, supporting multi-objective pedagogical goals

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

and highly personalized trajectories, albeit with increased data requirements and computational complexity [16, 17]. Crucially, AI-driven approaches do not universally outperform classical CAT; rather, classical methods remain essential for rigorous psychometric measurement, whereas modern algorithms extend adaptive selection toward richer behavioral contexts and multi-objective optimization. The resulting distinction between psychometric optimization, contextual exploration, and long-horizon policy learning provides a useful basis for evaluating adaptive selection methods across different educational assessment scenarios. Intelligent item selection determines which item should be delivered next, while the subsequent challenge is determining how sufficient, diverse, and appropriately targeted assessment content can be generated and evaluated at scale.

Table 2 Comparative Characteristics of Adaptive Item Selection Approaches

Approach	Core Mechanism	Scope	Key Trade-offs
Classical CAT	Iterative ability estimation	Myopic	Psychometric rigor vs. rigid item bank dependence.
Fisher Information	Maximizes local $I_i(\theta)$	Myopic	Local precision vs. item exposure imbalance.
Reinforcement Learning	MDP policy mapping	Long term	Multi-objective planning vs. high data demands.
Multi-Armed Bandits	Exploration-exploitation	Sequential	Cold-start adaptability vs. non-stationarity limits.

➢ *Automated Item Generation and Scoring*

- *Automated Item Generation*

The continuous delivery of adaptive assessments requires a sustained and scalable supply of high-quality test items, a demand that frequently overwhelms traditional manual authoring processes. To resolve this item-banking bottleneck, Automated Item Generation (AIG) has emerged as an essential methodology, shifting content creation from static repositories to dynamic, algorithmic generation [18]. Rather than relying solely on manual item writing by subject matter experts, AIG leverages computational methods to generate test questions systematically from specified learning objectives, source content corpora, underlying cognitive concepts, targeted difficulty levels, and learner capability models [18, 19].

Historically, item generation relied on template-based models, algorithmic item cloning, and psychometric rule-based frameworks where parameters were constrained by predefined syntactic templates. While these traditional techniques improved production efficiency, their structural rigidity limited linguistic diversity and contextual adaptability. Modern AIG paradigms extend these approaches by integrating advanced machine learning techniques capable of synthesizing novel assessment items while seeking consistency with specified content, difficulty, and psychometric constraints [18, 19].

The primary advantages of automated item generation include high operational scalability, enhanced content diversity, targeted personalization for specific student cohorts, and a substantial reduction in human authoring effort [18]. However, AIG systems encounter notable limitations. These include potential semantic ambiguity in generated stems, variable pedagogical quality, factual hallucinations, difficulty in precise a priori parameter calibration, and the persistent necessity for rigorous human validation prior to operational deployment [19]

- *Large Language Models for Assessment Content Creation*

The rapid evolution of transformer-based Large Language Models (LLMs) has fundamentally transformed automated assessment content creation. Instruction-tuned models are capable of generating question stems, comprehensive answer keys, detailed explanatory feedback, and contextually appropriate educational items across diverse domains [19]. These generative architectures leverage massive pre-trained semantic spaces to interpret source materials and produce tailored assessment content that aligns with specific pedagogical targets.

Within educational technology, generative AI applications—ranging from auxiliary tutoring tools to advanced assessment engines—have demonstrated considerable capacity to support educators by drafting formative checkpoints and diagnostic prompts [18, 19]. Nevertheless, it is critical to emphasize that LLM-generated assessment items are not automatically valid, reliable, or psychometrically equivalent to expert-authored items [19]. Studies evaluating generative outputs for higher education underscore that unverified model outputs frequently contain conceptual inaccuracies, biased phrasing, or misaligned difficulty levels [19], as further highlighted in empirical evaluations of LLM-generated multiple-choice questions [3]. Consequently, LLM-based content creation necessitates systematic validation protocols, expert review, and rigorous alignment with designated learning objectives to ensure assessment integrity.

- *Automated Distractor and Option Generation*

In multiple-choice assessments, the quality of incorrect alternatives—commonly known as distractors—directly dictates item discrimination power and overall test validity. Plausible distractors must challenge misconceptions without introducing ambiguity or providing superficial test-wiseness cues. Automated distractor and option generation addresses the complexity of authoring incorrect choices by leveraging instruction-tuned LLMs to synthesize contextually relevant alternatives alongside correct answers [19].

Modern generative frameworks utilize semantic conditioning to ensure that distractors target specific student errors, misconceptions, or partial knowledge states [19]. Despite these advancements, automated distractor generation introduces significant technical and psychometric risks. Common failure modes include the inadvertent generation of multiple correct answers, implausible distractors that fail to discriminate between proficiency levels, linguistic ambiguity, reliance on superficial lexical overlaps, and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

factual inaccuracies [19]. As systematic evaluations demonstrate, unverified generated options often require rigorous psychometric post-filtering and expert validation to prevent construct-irrelevant variance from compromising test scores [3].

- *Automated Scoring and Response Evaluation*

Once examinees complete the delivered assessment items, efficient and accurate evaluation is required to close the diagnostic feedback loop. Automated scoring systems streamline this process by transitioning beyond traditional binary grading toward sophisticated evaluation of constructed responses, encompassing automated short-answer scoring (ASAS) [21] and automated essay scoring (AES) [20]. Modern computational architectures rely heavily on transformer-based contextual embeddings and deep semantic representations to capture nuance, syntactic structure, discourse coherence, and conceptual alignment without depending solely on superficial keyword matching [20, 21].

Automated short-answer scoring evaluates concise student text against reference answers [21], whereas automated essay scoring analyzes extended compositions across multiple rhetorical and grammatical dimensions [20]. The operational benefits of automated scoring are substantial, encompassing high grading consistency, rapid feedback delivery, operational scalability for massive open online courses, and a significant reduction in evaluator workload [20, 21]. However, these systems exhibit notable limitations, including challenges in capturing deep semantic nuance, limited explainability in deep neural decisions, susceptibility to adversarial biases, domain dependency when transferring across distinct subjects, and occasional discrepancies with expert human raters [20, 21].

- *Quality, Validity, and Reliability Considerations*

The integration of generative artificial intelligence and automated scoring into high-stakes assessment environments introduces critical quality, validity, and reliability considerations. Generative AI and automated evaluation engines should not be portrayed as inherently superior to human educators; rather, they serve as computational augmentations that require rigorous oversight [22, 23, 24].

A central concern involves factual correctness and the mitigation of model hallucinations, where generated text appears fluent but contains erroneous pedagogical or domain-specific facts [19]. Furthermore, automated items must undergo strict difficulty calibration and psychometric evaluation to establish content validity, construct validity, and reliability across diverse demographic cohorts to prevent algorithmic bias and ensure equitable assessment outcomes [22, 24].

Because automated systems can occasionally produce opaque evaluations or biased scoring trajectories, model explainability is paramount [23]. Consequently, operationalizing AI-driven assessment requires robust human-in-the-loop validation frameworks, where expert educators audit generated content and verify automated scoring outputs before high-stakes deployment [22, 24].

- *Comparative Discussion*

Table 3 summarizes the principal characteristics, advantages, limitations, and validation requirements of the major automated item generation and scoring approaches.

A critical synthesis of content creation and evaluation methodologies reveals a clear technological evolution spanning traditional manual authoring, algorithmic item generation, Large Language Model-based synthesis, automated distractor generation, and automated response scoring. Traditional manual authoring offers high pedagogical reliability and psychometric control but suffers from prohibitive labor costs, restricted item pool size, and limited scalability. Automated item generation introduces computational efficiency and template-based scaling, yet remains constrained by rigid structural templates. The introduction of LLM-based generation and automated distractor design expands semantic diversity and personalization capabilities, though it introduces risks related to factual hallucination and validation overhead [18, 19, 3]. Finally, automated scoring transforms response evaluation by delivering rapid, consistent, and scalable analytics for complex constructed text, balancing computational efficiency against interpretability constraints [20, 21]. This progression demonstrates how modern educational technology has moved away from isolated, static content creation toward scalable, AI-assisted generation and evaluation ecosystems. However, generating and scoring assessment content is not sufficient by itself; these components must be seamlessly integrated with learner modeling, adaptive item selection, and overall assessment orchestration, providing the foundation for the subsequent sections of this survey.

➢ *System Architecture and Learning Analytics*

- *End-to-End System Architecture*

The practical execution of artificial intelligence-driven adaptive assessment systems requires an integrated computational architecture capable of coordinating the core pillars established throughout this survey: measurement models, adaptive item selection, automated item generation, and automated scoring [1, 2, 5, 6, 18, 20]. Rather than implying that any single existing software platform implements this exact comprehensive workflow, the architectural synthesis presented in this survey conceptualizes how these modular components interact within an integrated operational lifecycle.

The logical workflow progresses systematically: the learner interacts with the assessment interface, which captures response data and passes them to the response collection service; the learner-modeling engine updates the latent proficiency or cognitive state based on these inputs [5, 6]; the adaptive decision engine queries the item generation service or pre-calibrated item bank to determine the optimal next task [2, 16, 18]; upon completion, automated scoring modules evaluate constructed responses [20, 21]; and learning analytics engines process the resulting interaction logs to update instructor and learner dashboards, closing the feedback loop and refining subsequent iterations of the learner model [22, 24]. Within this framework, assessment

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

delivery services manage real-time session state, learner-modeling engines compute cognitive representations, item banks store psychometrically validated content, adaptive decision engines execute item-selection algorithms, automated generation services synthesize novel items or distractors [18, 19], scoring services process complex submissions [20, 21], and analytics engines aggregate performance metrics [24]. Instructor and learner dashboards visualize these insights to support pedagogical intervention and self-monitoring, all backed by secure data storage layers [24]

The overall end-to-end architecture integrating learner modeling, adaptive item selection, automated item generation, automated scoring, learning analytics, and continuous feedback is illustrated in Fig. 4.

- *Learner Modeling and Data Infrastructure*

The foundation of adaptive assessment relies upon robust data infrastructure capable of ingesting high-frequency interaction streams and transforming raw logs into meaningful learner representations [22, 23]. Core data elements typically include raw response histories, item identifiers, precise interaction timestamps, estimated proficiency parameters, latent knowledge states, skill mastery probabilities, multi-step interaction sequences, and aggregate assessment outcomes [1, 2, 5, 6]. Large-scale educational datasets and digital learning platforms provide the empirical volume necessary to train, evaluate, and benchmark these computational models [23].

Data infrastructure serves as the crucial bridge connecting the mathematical measurement models reviewed in Section IV with the adaptive selection mechanisms examined in Section V [1, 2, 5, 6, 16]. Maintaining data quality, temporal consistency, and strict provenance tracking is essential to prevent erroneous state updates during longitudinal tracking [22, 24]. Furthermore, handling missing interaction data and accurately representing dynamic learner states across sparse engagement logs present persistent technical challenges that require robust preprocessing pipelines within the data architecture [23].

Table 3 Comparative Analysis of Automated Item Generation and Scoring Approaches

Approach	Primary Function	Major Strengths	Key Limitations	Human Validation Requirement
1. Manual Item Authoring	Human creation of test items and rubrics	High pedagogical reliability and construct validity	High labor cost, limited item pool scalability	Complete expert review and pilot testing
2. Traditional AIG	Algorithmic generation via templates and rules	High production efficiency, consistent formatting	Structural rigidity, limited linguistic diversity	Psychometric calibration and expert audit
3. LLM-Based Generation	Neural synthesis of question stems and content	High scale, semantic flexibility, contextual variety	Factual hallucinations, variable difficulty calibration	Rigorous post-filtering and expert content validation
4. Distractor Generation	Creation of multiple-choice incorrect alternatives	Targets student misconceptions efficiently	Risk of multiple correct answers or implausible distractors	Psychometric review and distractor analysis
5. Automated Short-Answer Scoring	Evaluation of concise text responses	Rapid feedback, consistent grading of brief answers	Limited handling of deep semantic nuance	Periodic audit and exception handling
6. Automated Essay Scoring	Evaluation of complex constructed essays	Scalable analysis of discourse and structure	Domain dependency, limited deep structural explainability	Rater agreement auditing and bias monitoring

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

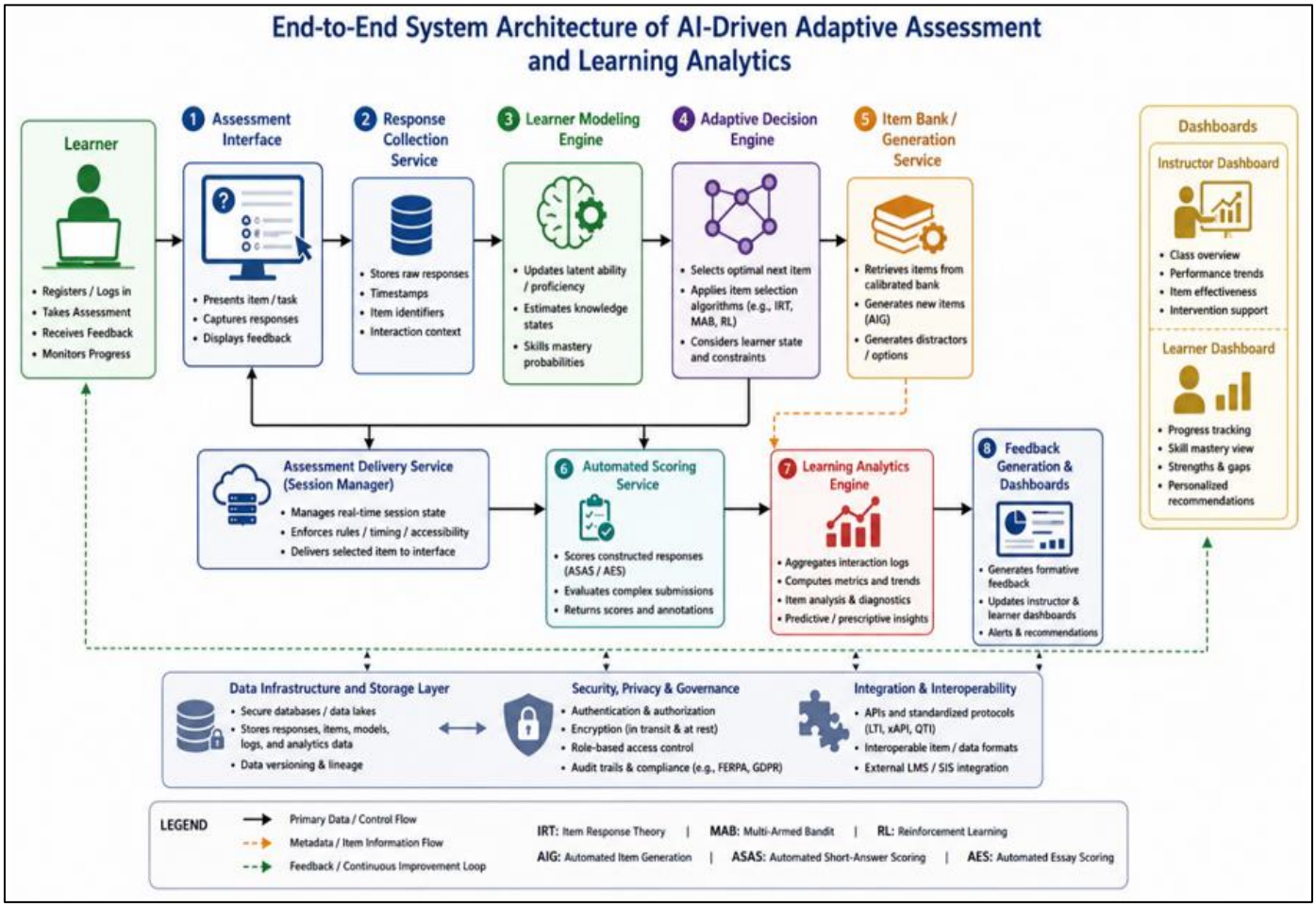

Fig 4 End-to-end System Architecture of Artificial Intelligence-Driven Adaptive Assessment, Integrating Learner Modeling, Adaptive item Selection, Automated item Generation, Automated Scoring, Learning Analytics, and Continuous Feedback.

- *Real-Time Assessment and Feedback Pipeline*

Real-time adaptive assessment depends upon an efficient operational loop that minimizes computational latency during live testing sessions [22, 23]. The execution cycle follows a continuous sequence: (1) delivery of the selected assessment item to the interface, (2) student interaction and response submission, (3) immediate capture of response data, (4) real-time updating of the learner state via measurement engines, (5) execution of adaptive item selection algorithms to identify the next optimal task, (6) invocation of automated scoring services where required, (7) dynamic generation of formative feedback, and (8) asynchronous updating of the learning analytics repository [5, 6, 18, 20, 22].

Achieving low-latency processing is critical to prevent testing fatigue and maintain an interactive testing experience [23]. Scalable implementations may employ distributed services or asynchronous processing to separate computationally intensive operations—such as neural network inference for automated scoring or large-scale item generation—from real-time session delivery [19, 21, 22, 23]. This separation ensures that performance bottlenecks do not disrupt the continuous feedback loop required to update learner representations dynamically [24].

- *Learning Analytics and Visualization*

Learning analytics transforms raw assessment interaction data into actionable educational information, bridging low-level system logs with high-level pedagogical insights [22, 24]. Analytic engines compute learner performance trends, longitudinal skill mastery trajectories, item difficulty parameters, response accuracy metrics, engagement indicators, item-level behavior, and assessment efficiency measures [22, 24]. These insights are delivered through dedicated instructor-facing dashboards and learner-facing visualizations designed to support formative intervention, self-monitoring, decision-making, and the identification of specific conceptual learning gaps [24].

When evaluating analytics outputs, it is critical to distinguish clearly between descriptive analytics—which summarize historical interaction data—and predictive or prescriptive analytics that forecast future performance or recommend adaptive learning pathways [22, 24]. Analytics dashboards offer valuable behavioral and diagnostic signals, but they should not be misinterpreted as definitive causal proof of cognitive learning gains without rigorous experimental validation [24]. As highlighted in educational data mining literature, behavioral evidence contributes valuable telemetry to learner modeling and personalized educational interventions, provided that descriptive metrics are properly contextualized.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

- *Scalability, Interoperability, Security, and Privacy*

The deployment of intelligent adaptive assessment systems in production environments introduces critical engineering challenges regarding scalability, distributed computing, and cloud-based infrastructure [22, 23]. Platforms must support large concurrent user populations during peak examination periods, requiring robust architectural design and elastic resource management [23]. Interoperability between heterogeneous educational platforms necessitates standardized data exchange protocols and portable model formats to ensure seamless integration across institutional boundaries [23].

Security and privacy are paramount because adaptive assessment systems continuously process sensitive learner interaction data, performance metrics, and behavioral logs [23, 24]. Robust authentication mechanisms, authorization controls, encrypted data storage, and privacy-preserving data management practices are essential to safeguard examinee information [23]. Protecting data provenance and maintaining compliance with institutional security standards remain critical prerequisites for deploying adaptive assessment platforms in high-stakes educational settings [24]. Furthermore, addressing algorithmic bias, ensuring fairness and equity across diverse demographic cohorts, and implementing explainable AI frameworks and human-in-the-loop oversight are vital to prevent discriminatory outcomes and maintain institutional trust [22, 24].

- *Comparative Discussion*

Table 4 summarizes system-architecture and learning-analytics approaches separately because these represent complementary rather than directly competing dimensions of an adaptive assessment platform.

Table 4 Comparative Analysis of System Architecture and Learning Analytics Approaches

Approach	Scalability	Processing / Output	Complexity & Limitation
PART A — SYSTEM ARCHITETURE			
Centralized	Low–Moderate	Low latency; limited integration	Low; small-scale use; peak-load limits
Distributed / Service-Oriented	High	Low latency; API-based integration	High; orchestration required
Cloud-Based / Hybrid	High (Elastic)	Low–Moderate latency; high interoperability	Moderate; cloud governance
PART B — LEARNING ANALYTICS			
Descriptive	High	Historical reports	Low; no prediction
Predictive	Moderate–High	Forecasts and risk indicators	High; sensitive to data drift
Prescriptive / Adaptive	Moderate–High	Adaptive recommendations	High; requires validation

A critical synthesis of architectural paradigms reveals distinct trade-offs across computational efficiency, flexibility, and deployment complexity. Centralized architectures offer simplicity and local data control but suffer from scalability bottlenecks under high concurrent loads [22]. Conversely, distributed service-oriented and cloud-based systems deliver high scalability, modular interoperability, and low-latency response times, albeit with increased infrastructure management complexity [22, 23]. Similarly, learning analytics methodologies range from descriptive reporting to predictive and prescriptive modeling, each balancing computational overhead against diagnostic utility [24].

Ultimately, system architecture and learning analytics provide the infrastructure through which the five taxonomy pillars operate as an integrated assessment ecosystem [22, 23, 24]. This integration ensures that theoretical psychometric models, adaptive item selection policies, generative content engines, and automated scoring modules operate harmoniously within scalable, real-world educational environments. Establishing this comprehensive architectural foundation sets the stage for examining the broader research challenges, limitations, fairness, explainability, privacy, deployment barriers, ethical considerations, and future research directions of artificial intelligence-driven adaptive assessment.

VI. CHALLENGES, LIMITATIONS, AND ETHICAL CONSIDERATIONS

➢ *Psychometric and Measurement Limitations*

Despite the computational advancements achieved by modern artificial intelligence-driven adaptive assessment systems, foundational psychometric and measurement limitations persist across both traditional and neural modeling paradigms [1, 2, 5, 6]. Classical Test Theory (CTT) remains constrained by sample and test dependence, making true score estimates vulnerable to population shifts [1]. Item Response Theory (IRT) addresses parameter invariance but relies heavily on strict mathematical assumptions, including unidimensionality, local independence, and accurate parametric model fit [2]. When these assumptions are violated in complex multi-skill environments, IRT parameter estimation can become unstable, resulting in biased ability estimates θ.

Longitudinal models also exhibit structural vulnerabilities. Bayesian Knowledge Tracing (BKT) assumes binary skill mastery and homogeneous learning rates across diverse student populations, frequently failing to capture rich inter-skill dependencies or variable learning trajectories [5]. Conversely, Deep Knowledge Tracing (DKT) and neural cognitive diagnosis models achieve high predictive accuracy on historical interaction logs but often lack transparent internal mechanisms, making it difficult to verify whether

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

high predictive performance reflects genuine cognitive acquisition or superficial pattern matching over temporal sequences [6]. Furthermore, greater predictive accuracy does not automatically imply better psychometric validity; optimizing neural loss functions for next-response prediction can conflict with core educational measurement goals such as diagnostic interpretability and construct representation. These limitations directly impact adaptive item selection, where distorted latent proficiency estimates can trigger suboptimal item routing, compounding measurement error over the course of an assessment.

- *Data Dependency and Generalization*

Modern AI-driven assessment systems depend heavily on large-scale, high-quality interaction datasets for training and calibration [22, 23]. However, real-world educational data are frequently characterized by sparse learner histories, missing response values, class imbalance, and significant measurement noise [23]. These data constraints introduce severe generalization barriers, including domain shift and distribution shift when models trained on specific institutional cohorts are deployed across different learner populations or academic domains [22, 23].

Furthermore, cold-start conditions—where new examinees or newly generated assessment items lack historical interaction records—pose substantial challenges for both machine learning classifiers and item selection engines [18, 19, 23]. Performance metrics demonstrated on benchmark educational datasets often fail to generalize automatically to new operational contexts without extensive local recalibration [23]. Consequently, data dependency limits the out-of-the-box portability of advanced assessment architectures, requiring continuous monitoring and data governance to maintain system reliability.

- *Algorithmic Bias and Fairness*

Algorithmic bias and fairness represent critical challenges in the deployment of AI-driven adaptive assessment systems [22, 24]. Sources of bias can originate from multiple stages of the assessment lifecycle, including training-data imbalances, underrepresented demographic cohorts, linguistic variations, socioeconomic disparities in digital access, historical educational inequities, and model-dependent item selection policies [22, 24, 25, 27].

Because adaptive systems tailor subsequent tasks based on evolving learner profiles, biased state representations or myopic item-selection criteria can systematically disadvantage specific student groups, restricting their access to appropriate difficulty levels or misestimating their true capabilities [16, 22, 24]. Addressing fairness requires viewing equity as an essential component of assessment quality rather than merely a post-hoc software engineering adjustment [24]. Without explicit fairness constraints and demographic parity audits, automated systems risk amplifying existing educational disparities under the guise of technological neutrality.

- *Explainability and Interpretability*

The trade-off between predictive flexibility and model explainability remains a central tension in intelligent assessment architecture [1, 2, 5, 6, 16, 19, 20, 21]. Classical models like CTT and IRT offer high mathematical transparency through explicit parameter definitions and item information curves [1, 2]. Similarly, BKT provides interpretable transition and guessing probabilities [5]. In contrast, advanced frameworks—such as recurrent neural networks in DKT [6], reinforcement learning policies for item selection [16], large language models for item generation [19], and deep neural architectures for automated scoring [20, 21]—often operate as black boxes.

Explainability is especially critical when AI systems influence high-stakes decisions, such as proficiency estimation, tailored item routing, automated scoring, and individualized educational recommendations [20, 21, 24]. When automated evaluations or item selections lack transparent rationales, educators and examinees cannot easily audit or contest algorithmic outputs [24, 26]. This opacity reduces institutional trust and complicates compliance with educational accountability standards, motivating the adoption of post-hoc interpretability methods and hybrid neural-psychometric architectures [5, 6].

- *Privacy, Security, and Data Governance*

The continuous collection of fine-grained learner interaction data—including response histories, behavioral traces, performance records, multi-step interaction sequences, and model-generated inferences—introduces substantial privacy and security risks [22, 23, 24]. Unlike static assessments that record only final scores, adaptive platforms ingest continuous telemetry streams that can expose sensitive cognitive and behavioral profiles [23, 24].

Privacy concerns center on unauthorized data profiling, intrusive tracking, and the potential exposure of personal learning trajectories, whereas security concerns involve vulnerabilities to unauthorized access, data breaches, and adversarial model manipulation [23]. Establishing rigorous data governance frameworks—encompassing secure data storage, strict access control, anonymization protocols, and transparent data retention policies—is essential to protect examinee rights and maintain institutional security within educational assessment ecosystems [23, 24].

- *Human Oversight and High-Stakes Deployment*

Given the technical limitations, bias risks, and opacity associated with automated algorithms, fully autonomous AI assessment remains problematic in high-stakes educational contexts [19, 20, 21, 22, 24]. Automated item generation can introduce factual hallucinations or misaligned difficulty levels [19], and automated scoring engines can occasionally diverge from expert human judgment [20, 21].

Consequently, operationalizing AI-driven assessment requires robust human-in-the-loop validation frameworks [22, 24]. Expert educators and psychometricians must audit generated items, review automated scoring decisions, calibrate difficulty parameters, monitor model drift, and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

provide accessible appeal and correction mechanisms for examinees [19, 20, 21, 24]. AI systems should function as computational augmentations that empower expert judgment rather than autonomously replacing human oversight in high-stakes testing environments [24].

➢ *Comparative Discussion*

Table 5 summarizes the principal challenges, their implications for adaptive assessment, and potential mitigation strategies.

Table 5 Challenges and Mitigation Strategies in AI-Driven Adaptive Assessment

Challenge	Risk	Impact	Mitigation
Psychometric	Assumption error	Biased estimates	Hybrid models
Data Shift	OOD error	Poor personalization	Domain adaptation
Bias	Demographic skew	Unequal outcomes	Fairness-aware models
Explainability	Black box	Low trust	XAI methods
Privacy	Data exposure	Privacy risk	Anonymization
Autonomy	AI errors	High-stakes risk	Human review

A critical synthesis of these challenges demonstrates that no single technical mitigation completely resolves the limitations inherent in intelligent educational assessment. Psychometric inaccuracies, data dependencies, algorithmic biases, explainability deficits, and privacy vulnerabilities interact across the assessment lifecycle [1, 6, 22, 24]. Consequently, building reliable, valid, and equitable adaptive assessment platforms requires coordinated, multi-layered mechanisms that integrate rigorous psychometric validation, transparent computational modeling, robust data governance, and active human oversight [19, 20, 22, 24]. Recognizing these multi-faceted limitations establishes the necessary foundation for exploring advanced research directions aimed at addressing open challenges.

These critical limitations motivate the need for stronger research directions involving robust multimodal learner modeling, trustworthy generative assessment, fairness-aware adaptation, explainable decision-making, privacy-preserving architectures, and human-centered validation, leading naturally into the exploration of future advancements.

VII. FUTURE RESEARCH DIRECTIONS

➢ *Multimodal and Continual Learner Modeling*

While current intelligent assessment systems rely predominantly on discrete response sequences and item identifiers [1, 2, 5, 6], future research could investigate multimodal and continual learner modeling. Emerging work in educational technology suggests that incorporating diverse behavioral telemetry—such as interaction timestamps, intermediate problem-solving steps, interface navigation patterns, and response latencies—could yield richer, temporally continuous student profiles. A promising research direction involves bridging the explicit interpretability of psychometric frameworks (such as Item Response Theory and Bayesian Knowledge Tracing) with the representational capacity of deep learning [1, 2, 5, 6]. However, integrating multimodal data does not automatically enhance assessment validity; future investigations must establish rigorous construct validation protocols to ensure that auxiliary behavioral signals measure intended cognitive competencies rather than construct-irrelevant variance.

➢ *Trustworthy Generative Assessment*

Building upon the item generation and scoring limitations examined in Section VI and Section VIII, future research in automated item generation must transition from unconstrained neural text synthesis to trustworthy, psychometrically controlled generation. Prospective investigations could focus on developing verifiable generation pipelines that integrate retrieval mechanisms, explicit constraint checking, automated quality filtering, and psychometrically guided difficulty calibration [18, 19]. Rather than treating large language models as autonomous replacements for expert item writers, an emerging research direction involves hybrid co-creation workflows where generative models assist educators by drafting targeted items, misconception-aware distractors, and diagnostic explanations that undergo strict expert review and empirical validation [19, 3].

➢ *Fairness-Aware and Explainable Adaptation*

Future development of artificial intelligence-driven adaptive assessment systems should move beyond optimizing predictive accuracy and measurement efficiency to explicitly incorporate algorithmic fairness, transparency, and accountability. Future work might explore real-time bias auditing frameworks, subgroup-aware adaptation policies, and continuous equity monitoring across diverse demographic cohorts [22, 24]. Furthermore, addressing the black-box nature of advanced neural learner models and reinforcement learning item-selection policies requires developing post-hoc interpretability tools and transparent architectures [1, 6, 16]. Providing explicit rationales for item routing, score computation, and adaptive recommendations represents an important research priority to ensure that examinees and educators can audit algorithmic outputs, thereby strengthening institutional trust and accountability [24].

➢ *Privacy-Preserving and Federated Assessment*

Because adaptive assessment platforms process continuous telemetry and sensitive behavioral profiles [22, 23, 24], future research should prioritize privacy-preserving methodologies. Prospective research directions include decentralized learning architectures, federated learning frameworks where student interaction models are updated locally without centralizing raw behavioral data, and differential privacy techniques for longitudinal learner

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

representation. While centralized cloud infrastructures currently dominate operational deployments, exploring decentralized, minimal-data assessment paradigms represents a crucial technical frontier for safeguarding examinee privacy and complying with data governance standards in educational technology [23, 24].

➢ *Long-Horizon Intelligent Assessment*

Extending beyond myopic item-selection strategies that optimize solely for local Fisher Information [2], future research should advance long-horizon adaptive assessment frameworks. Leveraging sequential decision-making and reinforcement learning principles [16, 17], future systems could investigate how to jointly optimize multiple educational objectives—including measurement precision, long-term cognitive gain, diagnostic coverage, learner engagement, item exposure control, and algorithmic fairness. A central research challenge involves designing reliable multi-objective reward functions that balance immediate testing efficiency with longitudinal learning trajectories, avoiding the assumption that reinforcement learning models universally outperform classical psychometric item selection [16, 17].

➢ *Human-Centered and High-Stakes Evaluation*

Future research must prioritize human-centered evaluation paradigms that position artificial intelligence tools as computational augmentations rather than autonomous decision-makers, particularly in high-stakes educational testing environments [19, 20, 21, 22, 24]. This entails developing robust expert-AI collaboration frameworks, transparent override mechanisms, continuous model-drift detection protocols, and standardized empirical methodologies for comparing automated outputs directly with expert human judgment [22, 24]. Emphasizing holistic system evaluation over isolated component accuracy will ensure that adaptive assessment platforms maintain operational safety, validity, and educational efficacy [24].

➢ *Research Priorities and Comparative Discussion*

A critical synthesis of these forward-looking pathways demonstrates that advancing intelligent adaptive assessment requires an interdisciplinary approach. Resolving current psychometric, algorithmic, and operational limitations demands coordinated progress across multimodal modeling, trustworthy generation, fairness auditing, privacy preservation, long-horizon optimization, and human-centered governance. These identified research directions point toward adaptive assessment systems that are more accurate, personalized, fair, explainable, privacy-aware, and human-centered, establishing the foundation for the final concluding summary of this survey. Table 6 summarizes the major research directions, their primary objectives, and key research focuses.

Table 6 Future Research Directions for AI-Driven Adaptive Assessment

Research Direction	Objective	Key Challenge	Research Focus
Multimodal Modeling	Integrate behavioral and cognitive telemetry	Construct validation and noise filtering	Telemetry integration
Trustworthy Generation	Verifiable, psychometrically aligned AIG	Hallucination control and calibration	Content reliability
Fairness & Explainability	Audit bias and provide transparent rationales	Balancing accuracy and transparency	Equity and auditing
Privacy-Preservation	Secure, decentralized learner profiling	Decentralized coordination overhead	Data governance
Long-Horizon Adaptation	Optimize multi-objective educational paths	Complex reward shaping and stability	Policy optimization
Human-AI Collaboration	Seamless expert audit and validation workflows	Operational overhead and workflow integration	Human oversight

VIII. CONCLUSION

This systematic survey has presented a comprehensive examination of artificial intelligence-driven adaptive assessment systems, establishing a unified end-to-end taxonomy that organizes the complete evaluation lifecycle. By structuring the literature across five core pillars—measurement models, adaptive item selection, automated item generation, automated scoring, and system architecture paired with learning analytics—this review bridges previously fragmented research directions into a cohesive operational framework.

The analysis reveals a deliberate technological evolution spanning traditional psychometric foundations, classical computerized adaptive testing, and modern neural representations, reinforcement learning algorithms, and generative content models. Crucially, the evidence demonstrates that artificial intelligence-driven approaches do not universally replace or outperform classical psychometric methods. Rather, classical psychometrics remains essential for measurement rigor, parameter invariance, and interpretability, whereas modern computational techniques extend adaptive personalization, sequential decision-making, content generation, and scalable response evaluation. While these enhanced paradigms offer substantial scalability and operational flexibility, they also introduce significant challenges concerning data dependency, algorithmic bias, model explainability, privacy vulnerabilities, and psychometric validity.

Ultimately, realizing the full potential of intelligent educational assessment requires harmonizing computational power with rigorous psychometric principles and active human oversight. Evaluating assessment quality demands a holistic system-level perspective that prioritizes validity,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

reliability, fairness, explainability, privacy, and human governance alongside predictive accuracy. By synthesizing current technological capabilities, critical limitations, and prospective research pathways, this survey provides a structured roadmap for researchers, educators, and system architects dedicated to developing trustworthy, scalable, and human-centered adaptive assessment ecosystems.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Dr. B. R. Mohan, Department of Computer Science and Engineering, East West Institute of Technology, Bengaluru, for his valuable guidance, constructive suggestions, and continuous support throughout the preparation of this survey paper. The authors also acknowledge the support and encouragement provided by the Department of Computer Science and Engineering, East West Institute of Technology.

REFERENCES

[1]. F. M. Lord, Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, 1980.
[2]. W. J. van der Linden and C. A. W. Glas, Elements of Adaptive Testing. Springer, 2010. doi: 10.1007/978-0-387-85461-8.
[3]. M. D. Reckase, Multidimensional Item Response Theory. Springer, 2009. doi: 10.1007/978-0-387-89976-3.
[4]. D. Magis and G. Raîche, "Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR," Journal of Statistical Software, vol. 48, no. 8, pp. 1–31, 2012. doi: 10.18637/jss.v048.i08.
[5]. A. T. Corbett and J. R. Anderson, "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge," User Modeling and User-Adapted Interaction, vol. 4, no. 4, pp. 253–278, 1994. doi: 10.1007/bf01099821
[6]. J. De La Torre, "The DINA Model Framework for Cognitive Diagnosis: Model Comparison and Invariance," Journal of Educational and Behavioral Statistics, vol. 34, no. 1, pp. 115–136, 2009. doi: 10.3102/1076998607309480.
[7]. N. Thai-Nghe, L. Drumond, T. Horváth, and L. Schmidt-Thieme, "Factorization Techniques for Predicting Student Performance," in Proceedings of the 6th ACM Conference on Recommender Systems (RecSys), 2012, pp. 129–136. doi: 10.1145/2365952.2365980.
[8]. C. Piech et al., "Deep Knowledge Tracing," in Advances in Neural Information Processing Systems (NeurIPS), vol. 28, 2015, pp. 505–513.
[9]. J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic Key-Value Memory Networks for Knowledge Tracing," in Proceedings of the 26th International Conference on World Wide Web (WWW), 2017, pp. 765–774. doi: 10.1145/3038912.3052631.
[10]. A. Ghosh N. Heffernan, and A. S. Lan, "Context-Aware Attentive Knowledge Tracing," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2330–2339. doi: 10.1145/3394486.3403272.
[11]. Q. Liu et al., "A Survey on Deep Knowledge Tracing: Models, Datasets, and Applications," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3341–3360, 2023. doi: 10.1109/TKDE.2021.3135248.
[12]. F. Wang et al., "Neural Cognitive Diagnosis for Intelligent Education Systems," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 615–623. doi: 10.1145/3394486.3403163.
[13]. A. Ghosh, N. Heffernan, and A. S. Lan, "Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable," in Proceedings of the 13th International Conference on Educational Data Mining (EDM), 2020, pp. 67–78.
[14]. W. Gao et al., "RCD: Relation-Constructed Cognitive Diagnosis for Intelligent Educational Systems," in Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM), 2021, pp. 524–533. doi: 10.1145/3459637.3482329.
[15]. F. Wang, Q. Liu, E. Chen, Z. Huang, and Y. Su, "Monotonic Neural Cognitive Diagnosis for Educational Assessment," IEEE Transactions on Learning Technologies, vol. 16, no. 2, pp. 210–224, 2023. doi: 10.1109/TLT.2023.3236056.
[16]. D. Nurakhmetov and A. S. Lan, "Reinforcement Learning Policies for Item Selection in Computerized Adaptive Testing," in Proceedings of the 15th International Conference on Educational Data Mining (EDM), 2022, pp. 112–123.
[17]. Y. Zhuang et al., "FCAT: A Fully Adaptive Computerized Testing Framework Based on Reinforcement Learning," ACM Transactions on Information Systems, vol. 41, no. 3, pp. 1–26, 2023. doi: 10.1145/3578532.
[18]. E. Kasneci et al., "ChatGPT for Good? On the Potential and Limitations of Large Language Models in Education," Computers and Education: Artificial Intelligence, vol. 4, p. 100135, 2023. doi: 10.1016/j.caeai.2023.100135.
[19]. A. Tack, C. Piech, and A. S. Lan, "Evaluating LLM-Generated Multiple Choice Questions for Higher Education Assessments," International Journal of Artificial Intelligence in Education, vol. 34, no. 1, pp. 145–172, 2024. doi: 10.1007/s40593-023-00360-1.
[20]. D. Ramesh and S. K. Sanampudi, "Automated Essay Scoring Using Transformer-Based Contextual Embeddings: A Comprehensive Review," Artificial Intelligence Review, vol. 55, no. 3, pp. 2495–2526, 2022. doi: 10.1007/s10462-021-10068-3.
[21]. E. Latif and X. Zhai, "AI-Based Automated Short Answer Scoring: Current Advances and Future Challenges," Educational Psychology Review, vol. 35, no. 2, p. 54, 2023. doi: 10.1007/s10648-023-09772-5.
[22]. M. Feng, N. Heffernan, and K. Koedinger, "Addressing the Assessment Challenge with an Online System that Tutors as It Assesses," User Modeling and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug610

User-Adapted Interaction, vol. 19, no. 3, pp. 243–266, 2009. doi: 10.1007/s11257-009-9061-0.

[23]. Y. Choi et al., "EdNet: A Large-Scale in-the-Wild Dataset for AI-Driven Education," in International Conference on Artificial Intelligence in Education (AIED), 2020, pp. 68–73. doi: 10.1007/978-3-030-52237-7_13.

[24]. Z. Wang et al., "NeurIPS 2020 Education Challenge: Diagnostic Questions Dataset," in NeurIPS 2020 Competition and Demonstration Track, 2021, pp. 125–140.

[25]. R. S. Baker and A. Hawn, "Algorithmic Bias in Education: Detection, Prevention, and Mitigation," WIREs Data Mining and Knowledge Discovery, vol. 11, no. 5, p. e1425, 2021. doi: 10.1002/widm.1425.

[26]. C. Conati, O. Barral, and K. Muldner, "Towards Explainable AI in Personalized Learning Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3-4, pp. 1–32, 2021. doi: 10.1145/3446340.

[27]. R. F. Kizilcec and H. Lee, "Fairness and Equity in Artificial Intelligence for Educational Assessment," Communications of the ACM, vol. 66, no. 7, pp. 82–91, 2023. doi: 10.1145/3579844.