

An AI-Driven Framework for Inclusive and Personalized Skill Assessment

Mr. Gaurav Kumbhar¹, Prof. Chetana M Kawale²

Research Scholar , Department of Computer Applications, SSBT's COET, Jalgaon, Maharashtra, India¹

Asst. Professor, Department of Computer Engineering, SSBT's COET, Jalgaon, Maharashtra, India²

Abstract: Assessment is a critical component of education, vocational training, and workforce skill development, yet conventional assessment systems continue to depend on fixed question sets, uniform response formats, and manual evaluation. These practices do not adequately account for differences in learner ability, preferred assessment mode, or accessibility requirements. This paper proposes and defines an evaluable AI-driven inclusive assessment framework that couples adaptive question selection, automated evaluation of objective and descriptive responses, accessibility support, skill-gap analysis, and performance analytics within a single platform, while retaining a human reviewer for uncertain or high-impact decisions. The framework supports multiple-choice, descriptive, practical, and viva-based assessment modes for candidates with a range of abilities and accessibility needs. The methodology follows nine interconnected stages, from requirement analysis and candidate modelling through adaptive selection, automated evaluation, accessibility support, skill-gap analysis, analytics, human review, and privacy governance. The enhanced pipeline can be compared, and evaluation is organised around measurable, component-wise criteria: question-selection efficiency and skill-estimation consistency for adaptive testing; scoring agreement (e.g., Cohen's kappa) and error analysis for automated evaluation; usability and accessibility-issue counts for the accessibility layer; and response-time and throughput for system performance.; instead, the exact baseline, metrics, and comparison procedure that a subsequent implementation will use are defined precisely. The principal contribution is not any single AI technique but the integration of previously separate capabilities into one accountable, measurable, and inclusive assessment workflow suitable for a diverse skill ecosystem.

Keywords: AI-Driven Assessment, Inclusive Assessment, Adaptive Testing, Automated Evaluation, NLP, Skill-Gap Analysis, Accessibility, Human-in-the-Loop

I. INTRODUCTION

Assessment plays an important role in determining whether learners have acquired the knowledge, skills, and competencies required for further education, vocational training, or employment. However, many conventional assessment systems rely on fixed question sets, standardized response formats, and manual evaluation. Such approaches can limit flexibility when candidates differ in prior knowledge, educational background, learning needs, assessment preferences, and accessibility requirements. These challenges become more significant in diverse skill ecosystems that include learners from schools, vocational institutes, training programmes, and different educational pathways. Inclusive assessment therefore requires more than simply digitizing conventional examinations. An effective assessment system should be able to accommodate different assessment modes, support learners with diverse accessibility needs, provide appropriate levels of difficulty, evaluate different types of responses, and generate meaningful information about learner performance and skill gaps. The problem considered in this study is consistent with the requirements identified in the Smart India Hackathon 2024 problem statement SIH1777, which emphasizes an AI-powered assessment solution supporting multiple assessment modes, accessibility, personalization, analytics, standardization, and data privacy [1]. In this research, these requirements are treated as a broader assessment-system problem rather than as a direct reproduction of the hackathon specification.

Existing research shows that individual AI techniques can address parts of this problem. Adaptive testing methods can select questions according to observed ability [4]; large language models can assist in scoring descriptive answers [2]; assistive technologies can improve accessibility for learners with disabilities [5][6]; and AI-based personalization can support equity and inclusion when applied carefully [3], provided generative-AI use follows recognised safeguards [7]. The research gap addressed in this paper is therefore the integration of these capabilities into a single, accountable, and experimentally testable assessment workflow. In particular, there is a need for a framework that connects adaptive question selection, automated response evaluation, accessibility support, skill-gap analysis, and performance analytics while maintaining human oversight for uncertain or high-impact decisions. Addressing this gap requires not only defining the individual components but also specifying how they interact and how their performance can be evaluated using measurable criteria.

Research Objectives

- O1. Design an adaptive question-selection mechanism and define a fixed-question-set baseline against which it can be compared.
- O2. Design a rubric-based automated evaluation process for objective and descriptive responses that retains human review for low-confidence or high-impact cases.
- O3. Integrate accessibility support and skill-gap analytics into the same assessment workflow rather than as disconnected add-on modules.
- O4. Define measurable, component-wise evaluation criteria (Section VII) that make the proposed framework experimentally testable in future implementation work.

II. LITERATURE REVIEW

The literature was selected to cover the five components addressed by the proposed framework: adaptive assessment, automated descriptive-answer evaluation, equity and inclusion, assistive technology for accessibility, and responsible governance of generative AI in education. Each study is summarised below in terms of its method, context, key finding, limitation, and relevance to the proposed framework, and Table 1 synthesises the six studies together.

Huang, Lin and Cheng (2009) — Adaptive Testing. A Computer Adaptive Testing (CAT) system combined with Item Response Theory (IRT) selects assessment items according to estimated learner ability within a versatile educational assessment platform. Adaptive item selection can maintain measurement precision while reducing the number of items a candidate must attempt, but the approach focuses on objective-item selection and does not address descriptive-response evaluation or accessibility. It forms the conceptual basis for the adaptive-assessment stage and the rule-based baseline defined in Section V [4].

Mughal, Imran, Daudpota, Kastrati and Noor (2026) — LLM-Based Automated Essay Scoring. Large language models were applied to automated essay scoring using benchmark datasets and classroom-collected data. LLMs can approximate human scoring reasonably well but remain sensitive to rubric design and can reproduce human rater bias, and the study recommends against using LLM scores as sole decision-makers. This supports the design choice, formalised in Section V, to treat automated descriptive scoring as an assistive rather than final decision [2].

Varsik and Vosberg (2024) — AI, Equity and Inclusion (OECD). This policy-oriented review examines cross-national AI applications in education. AI can support learner-centred personalization but can also introduce access, bias, and privacy risks that disproportionately affect already disadvantaged learners; the review is not tied to a specific assessment architecture and offers no experimental evaluation method. It motivates the fairness and privacy considerations built into Stage 9 of the proposed methodology (Section V) [3].

UNESCO IITE (2025) — Technologies for Inclusive Education. This multi-country review of assistive and ICT-based technologies for learners with disabilities and special educational needs finds that AI and assistive technologies, such as speech-to-text and text-to-speech, can materially improve access when properly integrated into the assessment interface, while identifying affordability and digital-literacy barriers that lie outside the scope of a single software framework. It directly informs the accessibility layer (Stage 5) and the accessibility metrics used in Section VII [5].

OECD (2026) — AI to Support Neurodivergent Learners in Vocational Education and Training. This policy review of AI tools deployed for neurodivergent learners in vocational settings finds that personalization and alternative interaction modes benefit neurodivergent learners specifically, but require careful evaluation to avoid stereotyping or misclassification, and it focuses on learning support rather than summative assessment. It reinforces that accessibility in the proposed framework must be evaluated as usability evidence rather than assumed from aggregate satisfaction scores (Section VII) [6].

Miao and Holmes (2023) — UNESCO Guidance for Generative AI in Education and Research. This policy guidance synthesises risks and safeguards for generative AI use in education, recommending explicit human oversight,

transparency about AI involvement, and safeguards against overreliance, though it does not propose a technical evaluation methodology. It directly informs the human-in-the-loop design (Stage 8) and the governance requirements described in Sections V and VII [7].

Ramesh and Sanampudi (2022)—Automated Essay Scoring: A Systematic Literature Review. This systematic review examines artificial intelligence and machine-learning approaches for automated essay and short-answer scoring. The authors discuss regression, classification, neural-network, and ontology-based approaches and identify continuing challenges in evaluating content relevance, coherence, and other dimensions of written responses. The study supports the use of automated descriptive-response evaluation in the proposed framework while highlighting the need for appropriate scoring criteria, validation, and human oversight for complex responses [8].

Xu, Mahmud, and Hoo (2024)—Automated Essay Scoring in Real-Life Education Scenarios. This systematic literature review examines automated essay-scoring systems reported between 2018 and 2023, with particular attention to their applicability in real educational environments. The review identifies challenges related to generalization across topics and writing styles, score reliability, robustness, and practical deployment. These findings support the proposed framework’s emphasis on reliability testing, error analysis, and validation rather than treating automated scores as universally reliable [9].

Idowu (2024)—Debiasing Education Algorithms. This systematic review examines fairness in machine-learning algorithms used in educational settings and discusses approaches for identifying and mitigating algorithmic bias. The review demonstrates that educational AI systems can introduce fairness concerns across different applications and therefore require explicit consideration of bias during system design and evaluation. These findings support the proposed framework’s inclusion of fairness considerations, human oversight, and component-wise evaluation as part of responsible AI-driven assessment [10].

Table 1. Synthesis of Reviewed Literature

Research Area	Representative Study	Method / Approach	Key Limitation	Use in Proposed Framework
Adaptive testing	Huang et al. [4]	CAT + IRT item selection	No descriptive scoring or accessibility support	Adaptive engine and rule-based baseline (Sec. V)
Automated scoring	Mughal et al. [2]	LLM-based essay scoring	Sensitive to rubric quality and rater bias	Rubric-driven descriptive evaluation (Sec. V, Stage 4)
Equity / inclusion	Varsik & Vosberg [3]	Cross-national policy review	No architecture or evaluation method	Fairness and privacy design goals (Sec. V, Stage 9)
Assistive technology	UNESCO IITE [5]	Multi-country technology review	Affordability and digital-literacy barriers	Accessibility layer and metrics (Sec. V Stage 5; Sec. VII)
Neurodivergent support	OECD [6]	Policy review, VET context	Focused on learning support, not assessment	Accessibility evaluated as usability evidence (Sec. VII)
Generative-AI governance	Miao & Holmes [7]	Policy guidance / safeguards	No technical evaluation methodology	Human-in-the-loop design (Sec. V, Stage 8)

III. PROPOSED FRAMEWORK

Building on the gap identified in Section III, this paper's core contribution is an integrated AI-assisted assessment framework, summarised in Fig. 1, rather than a single isolated AI model. The framework is organised as one pipeline: a candidate profile feeds an adaptive assessment engine; assessment responses pass to an answer-evaluation layer that is NLP/LLM-assisted for descriptive items; evaluation evidence feeds skill-gap analysis and performance analytics in parallel; these are combined into personalized feedback; and the entire chain is subject to human verification before a final assessment outcome is issued.

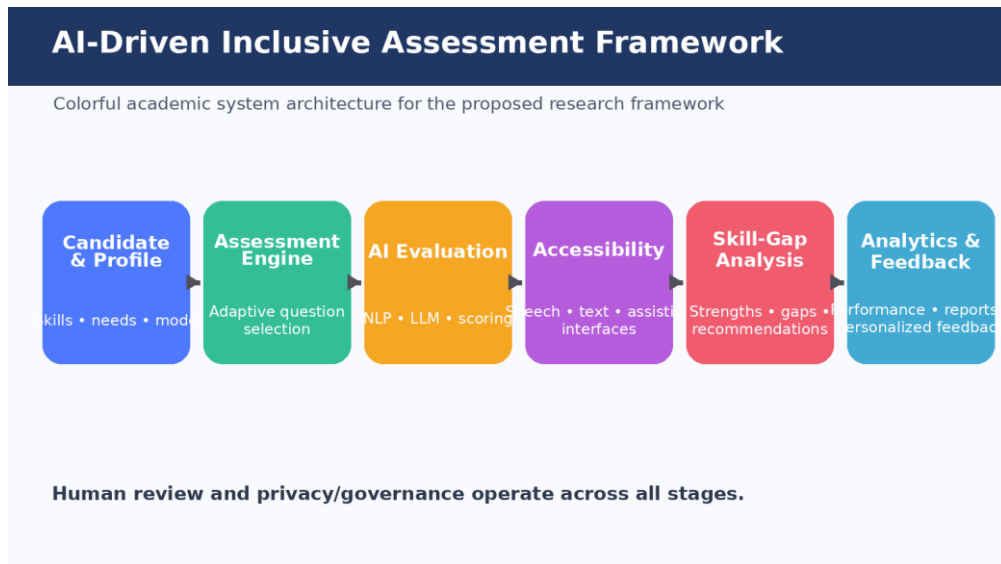


Fig. 1. Proposed AI-Driven Inclusive Assessment Framework

Two design decisions distinguish this contribution from the studies reviewed in Section II. First, automated assistance is explicitly separated from final decision-making: objective responses are processed automatically, but descriptive responses are evaluated against an explicit rubric and flagged for human review whenever confidence is low or the decision carries significant consequence for the candidate. Second, the framework treats the assessment score as an intermediate result rather than a final output: scores are mapped to skill or competency categories, producing skill-gap indicators, strengths, and next-step learning recommendations. This second decision connects assessment to learning, extending the framework's relevance beyond a single examination event to continuous skill development within an institution.

IV. METHODOLOGY

The research follows an applied, experimental framework-design approach. The methodology is organised into nine interconnected stages (Table 3) so that personalization, evaluation, accessibility, analytics, and human oversight operate within one assessment workflow rather than as disconnected modules.

Table 3. Proposed Methodology Stages

Stage	Methodology Stage	Description
1	Requirement Analysis and Assessment Design	Define candidate profiles, skill domains, question types, scoring rubrics, accessibility needs, and supported modes; question-bank metadata includes skill, topic, difficulty, marks, question type, and expected competency.
2	Candidate and Question Modelling	Represent candidate performance through prior responses, scores, and competency estimates; represent question items with difficulty and skill attributes to support adaptive selection.
3	Adaptive Assessment Engine	Select the next question based on observed performance, beginning with rule-based difficulty adjustment (the baseline in Section 5.1) and later evaluating formal CAT/IRT methods.
4	Automated Answer Evaluation	Score objective responses directly; process descriptive responses with NLP features and, where appropriate, a language model, against an explicit rubric, with human review for uncertain cases.

Stage	Methodology Stage	Description
5	Accessibility Layer	Provide speech-to-text, text-to-speech, keyboard accessibility, readable presentation, and timing accommodations appropriate to the candidate and assessment mode.
6	Skill-Gap Analysis	Map candidate responses to skill/competency categories to identify strengths, weak areas, topics requiring revision, and suggested next learning actions.
7	Performance Analytics	Present score trends, question-level and skill-wise performance, completion and response-time indicators, avoiding unsupported AI inferences.
8	Human Review and Decision Support	Flag uncertain evaluations and high-impact cases for reviewer inspection; the final academic or administrative decision remains with an authorized human reviewer.
9	Privacy and Governance	Collect candidate data only for legitimate assessment purposes, protect it through access controls, and log model outputs and review events for audit.

Proposed Technology Stack

A prototype may use Python for the backend and machine-learning layer; Pandas and NumPy for data processing; Scikit-learn for baseline models; NLP libraries such as NLTK or spaCy; a database such as PostgreSQL or SQLite; and Flask/FastAPI for a service layer. A web interface can provide the candidate assessment view and an evaluator dashboard.

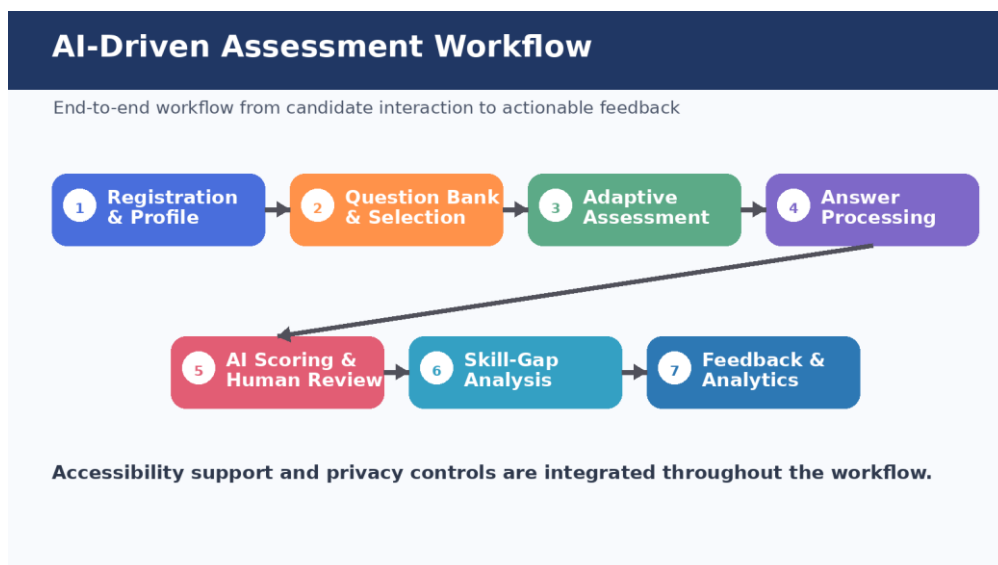


Fig. 2. AI-Driven Assessment Workflow

V. SYSTEM ARCHITECTURE AND WORKFLOW

The proposed architecture is a sequence of connected layers. A data layer manages candidate profiles, question-bank metadata, assessment rules, and accessibility requirements. A modelling layer represents candidate performance and question characteristics (Stage 2). The adaptive-assessment layer selects the next question based on observed performance (Stage 3). The response-evaluation layer processes objective and descriptive answers (Stage 4). The resulting evidence is passed to skill-gap analysis and performance analytics (Stages 6–7), and a human-review layer handles uncertain or high-impact cases (Stage 8), with privacy governance (Stage 9) applied throughout rather than as a final step.

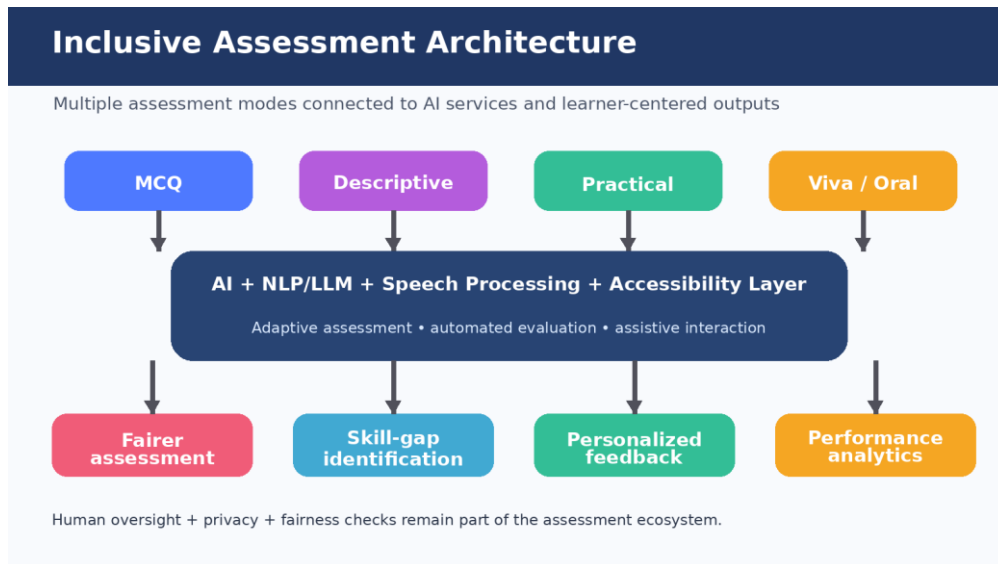


Fig. 3. Inclusive Assessment Architecture

The assessment workflow supports multiple modes, including multiple-choice questions, descriptive responses, practical tasks, and viva-based assessment, so that the same framework maintains common assessment objectives while allowing different response mechanisms. Accessibility features described in Stage 5 are attached at the interface layer rather than duplicated per mode, which keeps the accessibility implementation consistent across MCQ, descriptive, practical, and viva formats.

This staged, layered structure allows the adaptive engine, evaluation layer, and analytics layer to be implemented, tested, and replaced independently as long as they honour the same interfaces — for example, a rule-based adaptive engine can later be swapped for a formal CAT/IRT implementation without requiring changes to the evaluation or analytics layers.

VI. DESIGN AND EVALUATION PLAN

Dataset and Annotation

A prototype evaluation requires a dataset containing question metadata and candidate responses. For supervised components, labelled examples should be prepared using documented annotation rules; where real candidate data are unavailable, a controlled prototype dataset may be used with its limitations clearly stated.

Adaptive Testing Evaluation

The adaptive configuration (Table 4) is compared against the fixed-set baseline on question-selection efficiency (number of items needed to reach a stable skill estimate), skill-estimation consistency across repeated administrations, and assessment completion time.

Automated Scoring Evaluation

Automated scores are compared with human-assigned rubric scores using Cohen's kappa or a similar agreement statistic, accuracy where discrete labels permit, and precision/recall/F1 for any classification-style sub-task (for example, pass/fail rubric criteria). Particular attention is paid to error analysis of cases where the automated score is unexpectedly high or low relative to the human rubric score.

Accessibility and Usability Evaluation

Candidate feedback is collected on ease of use, clarity, accessibility, confidence, and perceived fairness, with testing across different interaction requirements where feasible. Accessibility is evaluated as a system property, measured through usability, task completion, and accessibility-issue counts, rather than inferred only from overall candidate satisfaction.

System Performance Evaluation

System response time, assessment completion time, and processing overhead are recorded. For larger deployments, database indexing, caching, model selection, and efficient candidate retrieval become important engineering considerations and are tracked as throughput and processing-time metrics.

Fairness, Reliability and Privacy Evaluation

Fairness and reliability require explicit error analysis: the evaluation inspects unexpectedly high or low scores and, where data quality and ethical approvals permit, examines outcomes across relevant candidate groups. Reviewer override rates indicate where automated outputs require frequent correction. Privacy and governance (Stage 9) are audited by confirming that candidate data are collected only for legitimate assessment purposes, protected by access controls, and that model outputs and review events are logged sufficiently for later audit. These measurements are interpreted carefully because the quality and representativeness of the assessment dataset directly influence the reliability of the findings.

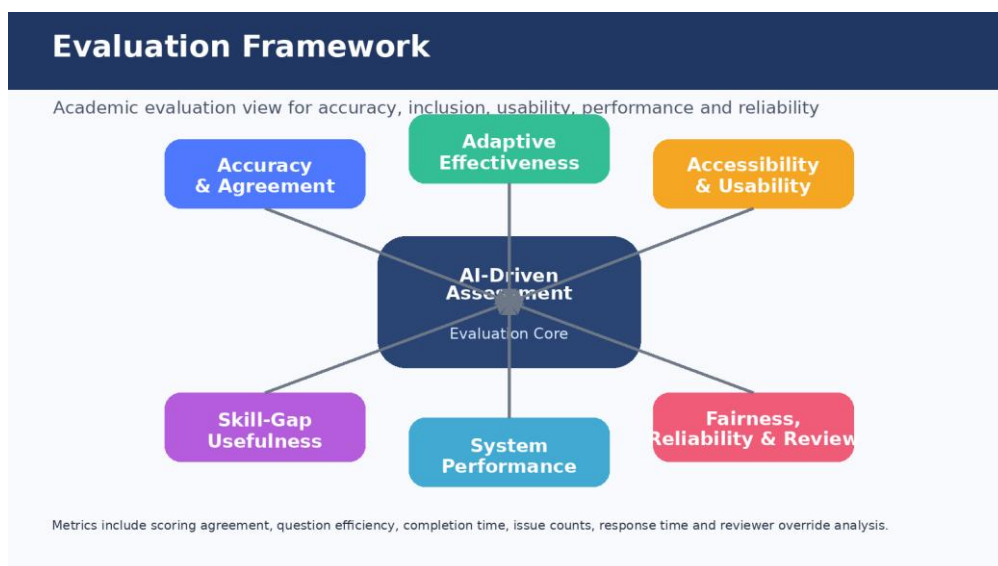


Fig. 4. Evaluation Framework

VII.RESULTS

Because this paper presents a proposed framework rather than completed experimental measurements, numerical performance values are not claimed in advance. The expected results describe what the prototype should produce, relative to the baseline in Table 4, once implemented and tested.

a. Personalized Assessment. Relative to the fixed-set baseline, the adaptive configuration is expected to select questions according to observed candidate performance, giving more challenging items to strong performers and suitable foundational items to weaker performers.

b. Automated Descriptive Evaluation. The proposed configuration is expected to assist in evaluating descriptive answers using the shared rubric and NLP/LLM techniques, producing a score or evaluation explanation reviewable by a human evaluator, with higher rubric agreement than the baseline's keyword-matching approach.

c. Skill-Gap Identification. Candidate responses are expected to be mapped to skill categories, allowing the system to identify stronger competencies and areas requiring improvement — a capability absent from the baseline.

d. Personalized Feedback. The system is expected to generate understandable, learning-oriented recommendations based on observed performance rather than only a final score.

e. Accessibility Support. The platform is expected to provide alternative interaction methods where appropriate, enabling candidates with different accessibility needs to participate more effectively than under the baseline.

f. Performance Analytics. Evaluators are expected to receive dashboards containing relevant performance indicators, such as scores, skill-wise results, response time, and flagged cases.

g. Human-in-the-Loop Verification. Cases with low confidence, ambiguous responses, or important consequences should be routed to a human reviewer, reducing the risk of blindly accepting an AI-generated assessment outcome.

VIII. DISCUSSION

Returning to the research questions in Section I, the framework proposed in Sections IV–VI directly addresses Q1–Q4: it treats adaptivity, evaluation, accessibility, and skill analytics as stages of one pipeline (Q1); it improves on siloed prior work by sharing one rubric and dataset across baseline and proposed configurations (Q2); it fills the integration gap identified in Section III (Q3); and its contribution is the accountable, measurable integration itself rather than any single AI technique (Q4). A central design issue is the balance between automation and reliability. More automation can reduce evaluator workload and provide immediate feedback, but it can also amplify model errors or bias, which is why the system distinguishes routine, high-confidence cases from uncertain cases that require human attention (Objective O2). A system is not inclusive merely because it uses AI. Candidates may differ in language, digital literacy, physical ability, neurodiversity, device access, and preferred interaction method, so the platform must provide appropriate alternatives while maintaining equivalent assessment objectives — which is why accessibility is measured as its own system property in Section VII rather than assumed (Objective O3).

The framework also connects assessment with learning. Traditional assessment often ends with a score, whereas the proposed system translates performance into skill-level indicators, identified gaps, and personalized feedback, which is more useful for skill-development institutions focused on continuous improvement than a single pass/fail outcome. Scalability is a further consideration: a large skill ecosystem may contain many candidates, questions, and assessment records, so efficient database design, indexing, caching, and appropriate model selection are required for real-time or near-real-time analytics. Advanced language models may provide richer semantic evaluation but can increase computational cost, latency, and governance requirements, which is why Stage 9 treats governance as a continuous requirement rather than a one-time setup step.

IX. CONCLUSION

This paper proposed an AI-driven inclusive assessment framework for a diverse skill ecosystem that integrates adaptive assessment, automated answer evaluation, accessibility support, skill-gap analysis, personalized feedback, performance analytics, human review, and privacy-aware governance within one accountable workflow. The literature review showed that these components have meaningful individual potential but are typically studied in isolation, motivating the specific integration gap this paper addresses (Section III). Unlike a purely descriptive proposal, this paper defines an explicit baseline configuration (Table 4) and a component-wise evaluation plan with named metrics (Table 5), so that the framework's practical value can be measured rather than assumed once implemented. The framework is deliberately positioned as decision support rather than a replacement for authorized evaluators, because automated assessment can introduce errors, bias, privacy concerns, and reliability issues that require systematic measurement and continued human oversight.

The next stage of this research is empirical implementation and testing of the baseline and proposed configurations. By comparing them on question-selection efficiency, scoring agreement, accessibility and usability evidence, system performance, and fairness-related error analysis, future work can establish which components of the proposed framework provide practical value for inclusive skill assessment, and which require further refinement before deployment.

REFERENCES

- [1] Smart India Hackathon. (2024). AI-Driven Inclusive Assessment Tools for Skill Ecosystem (SIH1777), Software Problem Statement. Smart India Hackathon 2024.
- [2] Mughal, N., Imran, A. S., Daudpota, S. M., Kastrati, Z., & Noor, W. (2026). Exploring potential of large language models for automated essay scoring in education. *Discover Artificial Intelligence*, 6, Article 166. <https://doi.org/10.1007/s44163-026-01002-y>
- [3] Varsik, S., & Vosberg, L. (2024). The potential impact of Artificial Intelligence on equity and inclusion in education. *OECD Artificial Intelligence Papers*, No. 23. OECD Publishing. <https://doi.org/10.1787/15df715b-en>
- [4] Huang, Y.-M., Lin, Y.-T., & Cheng, S.-C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers & Education*, 52(1), 53–67. <https://doi.org/10.1016/j.compedu.2008.06.007>
- [5] UNESCO Institute for Information Technologies in Education. (2025). *Innovative Technologies for Inclusive Education: A Review of Best Practices from Global Resource Centers*. UNESCO IITE.
- [6] OECD. (2026). *AI to Support Neurodivergent Learners in Vocational Education and Training*. OECD Publishing, Paris. <https://doi.org/10.1787/718d7522-en>
- [7] Miao, F., & Holmes, W. (2023). *Guidance for Generative AI in Education and Research*. UNESCO, Paris.
- [8] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- [9] Xu, W., Mahmud, R., & Hoo, W. L. (2024). A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? *IEEE Access*, 12, 77639–77657. <https://doi.org/10.1109/ACCESS.2024.3399163>
- [10] Idowu, J. A. (2024). Debiasing education algorithms. *International Journal of Artificial Intelligence in Education*, 34, 1510–1540. <https://doi.org/10.1007/s40593-023-00389-4>