International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

# ML-DRIVEN DYNAMIC TEST ITEM GENERATION

**Lankepalli. Kaveri#1, Kavali. Silpa#2, Kanaka. Ganayathri#3 , MR.P. Jayachandran #4**

**Email :, kaverilankepalli@gmail.com, kavalisilpa9@gmail.com, ganakanaka@gmail.com**

**#123 UG Student, Department of Computer Science and Engineering, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Trichy-621112-Tamilnadu.**

**#4 Assistant Professor, Department of Computer Science and Engineering , Dhanalakshmi Srinivasan Institute of Technology, Trichy-621112- Tamil Nadu.**

***Abstract-*** *Machine Learning–driven dynamic test item generation has gained significant attention in modern educational and assessment systems due to the increasing demand for scalable, adaptive, and personalized evaluation methods. The primary objective of such systems is to automatically generate high-quality test items that align with learner proficiency levels while reducing the manual effort involved in traditional question paper design. This project investigates the application of machine learning and natural language processing techniques to analyze educational content and generate contextually relevant, difficulty-adaptive test items. The proposed approach focuses on leveraging data-driven models to understand semantic relationships, learning objectives, and cognitive complexity within instructional material. Limitations of existing manual and rule-based test generation methods, including lack of adaptability, time inefficiency, and limited personalization, are identified through a review of current assessment practices. The findings emphasize the potential of ML-based dynamic test item generation systems to enhance assessment accuracy, support intelligent learning platforms, and improve the overall effectiveness of technology-driven.*

**Keywords -**
ML, Natural language Processing, Dynamic test generation, Adaptive assessment.

## I. INTRODUCTION

Dynamic Test Item Generation (DTIG) systems act as intelligent assessment frameworks that support effective learning evaluation, adaptive testing, and educational quality management. These systems are designed to ensure that assessment items are generated accurately, contextually aligned with learning objectives, and tailored to varying learner proficiency levels while minimizing redundancy and content irrelevance. With the rapid growth of digital learning platforms and large-scale online education, the demand for intelligent and automated test generation has become a critical requirement in modern educational environments. ML-driven DTIG systems aim to address these challenges by automatically generating relevant and difficulty-adaptive test items from educational content using machine learning and natural language processing techniques. Enhanced generation capabilities assist educators and learners by reducing manual question design efforts, enabling real. Fundamentally, ML-driven DTIG systems exhibit the following key characteristics that distinguish them from traditional assessment approaches:

- Content Information Extraction, which involves identifying relevant Key Concepts.
- Test Item classification and organization, which focus on items.

The systems play a crucial role in adaptive learning environments by facilitating personalized assessment strategies tailored to individual learner needs. These systems are developed to automatically generate test items that reflect content difficulty, cognitive level, and learning objectives. As digital education systems continue to expand, traditional manual assessment creation methods face limitations in scalability and responsiveness. ML-driven DTIG systems address these challenges by utilizing intelligent algorithms to extract semantic meaning from educational content and dynamically generate assessment items. Such systems enhance learning outcomes by providing timely feedback, enabling continuous skill evaluation, and improving learner engagement through adaptive learning.

## II DTIG DEVELOPMENT FRAMEWORK

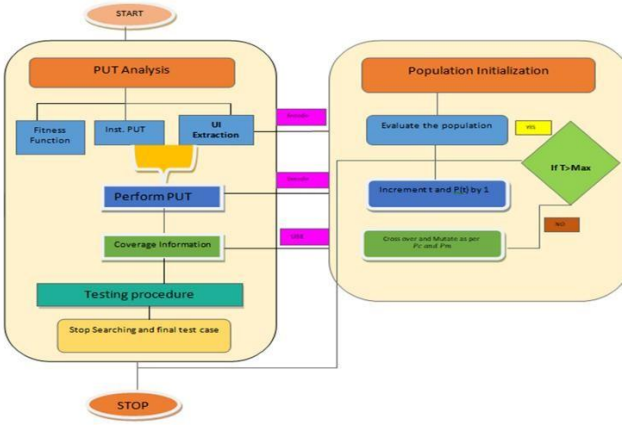



**Fig.1 DTIG Development**

The functional and computational components of Dynamic Test Item Generation (DTIG) systems are tightly interconnected, as the effectiveness of generated test items depends on both the underlying data processing mechanisms and the system architecture. Consequently, the capabilities of a DTIG system are determined by its core components, including data sources, machine learning models, content representation techniques, and evaluation mechanisms. The operational modes of the DTIG system, such as offline batch generation and real-time adaptive assessment, significantly influence its overall design and deployment strategy. In order to balance model complexity, computational efficiency, and content quality, DTIG systems must be carefully designed to integrate automated generation with validation and refinement processes. Conceptually, DTIG systems can be positioned between two extremes: a basic generation framework, where test items are produced from predefined templates and manually reviewed, and an advanced intelligent DTIG approach, where items are dynamically generated, difficulty-calibrated, and evaluated automatically with minimal human intervention. Another critical aspect is system robustness, particularly in high-stakes assessment environments where incorrect or biased test items can negatively impact learning outcomes. Testing DTIG systems presents significant challenges, as it is impractical to anticipate all variations in input content, learner behavior, and contextual constraints during evaluation phases. This paper is structured as follows: Sections II and III review existing industrial and academic methodologies for automated test item generation, while Section IV discusses emerging challenges in adaptive and large-scale assessment systems that demand increasingly intelligent, scalable, and reliable DTIG solutions. .

## III. THE CURRENT PROCESS OF DTIG DEVELOPMENT AND INDUSTRIAL PRACTICES

This section outlines the procedures currently employed in academic and industrial environments for the design, development, validation, and deployment of Dynamic Test Item Generation (DTIG) systems. DTIG system development must be viewed as a comprehensive, system-level process that requires a thorough understanding of educational assessment theory machine learning model design, and software engineering principles.

It was established by the Educational Technology for Standardization constitute the basis for DTIG development practices. The many stages of the DTIG system development and how they relate to the various stages of the Content analysis project are depicted

### A. DTIG's INITIAL LIFE CYCLE STAGE

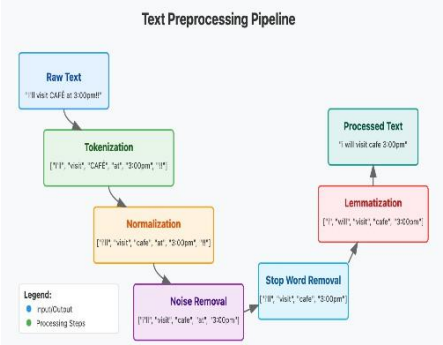
The requirements and core functionalities of a Machine Learning–Driven Dynamic Test Item Generation (DTIG) system are defined during the initial life cycle stage based on high-level educational objectives, assessment constraints, curriculum standards, and the outcomes of early content and risk analysis processes. Analytical techniques analogous to Failure Modes and Effects Analysis (FMEA) and dependency analysis are employed to identify potential risks related to content ambiguity, difficulty misalignment, bias, and question validity within automated item generation workflows. Deductive analysis focuses on high-impact assessment failures, such as incorrect difficulty classification, concept misrepresentation, or misleading answer options, and traces their root causes across data sources, feature extraction methods, and model architectures. Conversely, inductive analysis evaluates how specific issues—such as incomplete instructional content, noisy datasets, or model generalization errors—may influence overall item quality, fairness, and assessment reliability. Once the functional requirements of the DTIG system are established, the corresponding software components, including content preprocessing modules, machine learning models, difficulty calibration mechanisms, and item validation engines, are developed and integrated into the learning or assessment platform..

### B. ARCHITECTURE OF THE DTIG SYSTEM

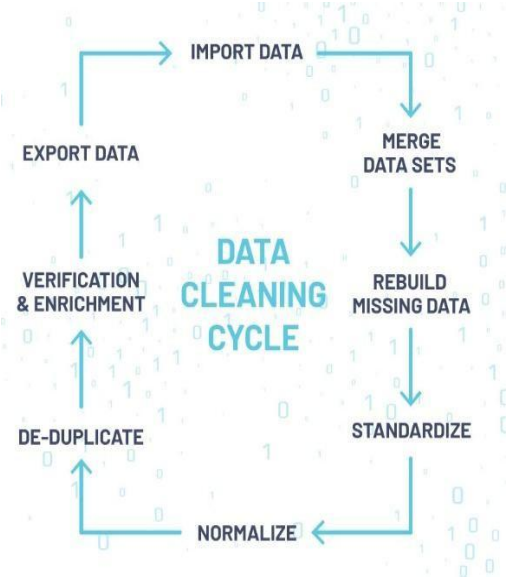
The DTIG system architecture is typically structured into multiple hierarchical layers, each defined by distinct interfaces, levels of abstraction, processing responsibilities, and output granularity. The lower layers, which are primarily data-driven and software-oriented, operate at the content acquisition and preprocessing level and are responsible for tasks such as instructional content ingestion, text normalization, concept extraction, and data validation to ensure content consistency and relevance. When these lower layers are unable to adequately resolve challenges such as semantic ambiguity, incomplete learning material, or noisy input data, higher-level layers are invoked to perform contextual interpretation and pedagogical analysis. During the system definition and early development phases, DTIG concepts are refined and translated into a comprehensive system architecture, with clear identification and allocation of functional responsibilities across generation, evaluation, and adaptation components. The architecture is designed to support efficient content processing, reliable test item generation, and scalable assessment delivery while maintaining fairness, interpretability, and integration with existing learning platforms. In this framework, higher architectural layers leverage the processed outputs, metadata, and confidence indicators produced by lower layers to perform advanced machine learning inference, difficulty calibration, and item validation..

***The following levels make up the hierarchical structure of the DTIG system:***

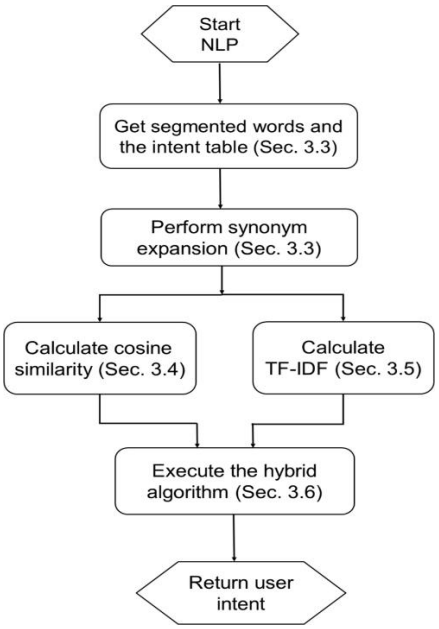
- ***Level 0:*** addresses data noise, formatting inconsistencies, tokenization errors, missing metadata, and other faults that are local to individual content units



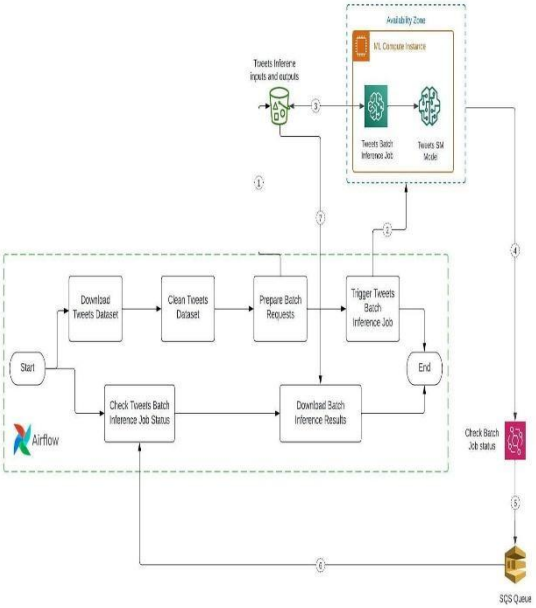

- ***Level 1***: manages content inconsistencies and errors that extend beyond a single data source and operate at the module or subsystem level. Examples include conflicting learning objectives across materials, incomplete instructional content, duplicated concepts, or mismatched difficulty labels..



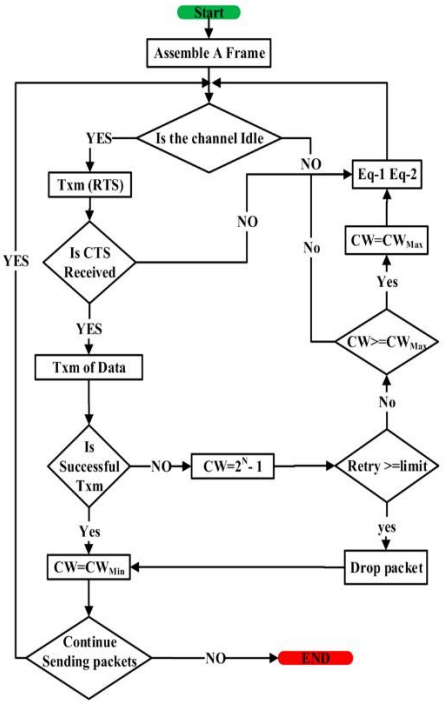

- ***Level 2:*** Level 2 handles failures associated with system-level test item generation, such as incorrect concept interpretation, poor distractor generation, difficulty misclassification, or loss of contextual coherence. Detection mechanisms rely on semantic validation, cross-content analysis, and pedagogical consistency checks. Failures at this level may result in partial degradation of item quality or assessment reliability.




- ***Level 3:*** Level 3 is associated with failures in the central AI processing components, including NLP models, item generation engines, difficulty calibration modules, and inference pipelines. Fault management strategies include model confidence scoring, performance monitoring, fallback to template-based or rule-driven generation, controlled degradation of functionality, and system reinitialization.



- ***Level 4:*** Level 4 encompasses critical failures that may compromise the integrity, fairness, or availability of the assessment system. These include widespread model bias, large-scale generation faults, systemic content corruption, or failures affecting multiple assessment subsystems simultaneously. The primary requirement at this level is complete functional independence and redundancy between the nominal DTIG components and the monitoring and recovery mechanisms. Recovery strategies may involve isolating affected subsystems, reverting to validated question banks, halting dynamic generation, and enabling manual oversight until corrective actions are completed.




In order to achieve a more economical and scalable implementation, a common trend in modern educational technology systems is to shift several traditionally manual or rule-based assessment functions into software-driven intelligent modules In this context, DTIG functionality is increasingly realized through machine learning–based software components, as this approach has become a standard practice in data-driven learning platforms. By embedding test item generation logic within adaptive software frameworks.

configurations, which stand for the system level DTIG's final response to serious abnormalities in the Test. Both ground and urgent events occurring on board can put the Test in safe mode. For a predetermined amount of time, the spacecraft can operate in this mode without assistance from the ground segment.

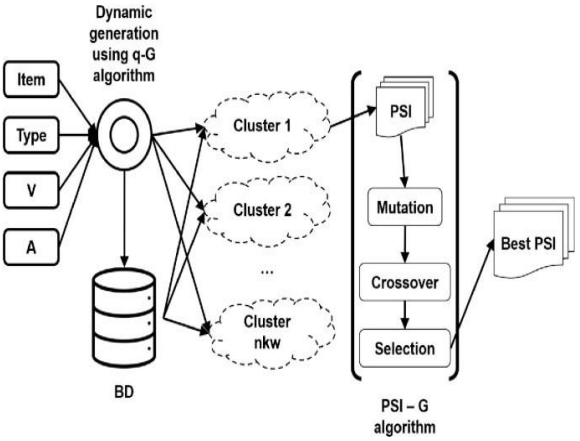



**Fig.2 DTIG Hierarchical structure**

In critical operational conditions, the DTIG system transitions into a controlled fallback mode in which automated test item generation capabilities are restricted, non-essential processing modules are suspended, and access to validated assessment content is preserved. During this mode, core safeguards such as content integrity checks, bias monitoring, logging mechanisms, and audit trails remain active to ensure assessment reliability and fairness. Human oversight by educators, assessment designers, or system administrators is required to restore the DTIG system from fallback mode to full operational status after verification of content consistency, model performance, and generation accuracy. During the development of software requirements, DTIG functional requirements and associated software artifacts undergo a structured Verification and Validation (V&V) process aligned with established educational software engineering and AI system validation practices. Software assurance levels are defined based on the potential impact of anomalous system behavior, ranging from Level A (where failures could critically compromise assessment validity .

Accordance to ensure continuous and reliable operation, DTIG systems incorporate monitoring and control mechanisms that dynamically assess system health, content quality, and model performance during runtime. These mechanisms track indicators such as confidence scores, item validity metrics, bias indicators, and learner response patterns to detect deviations from acceptable operating conditions. When predefined thresholds are exceeded, the system automatically triggers protective actions, including throttling generation rates, isolating affected modules, or switching to conservative generation strategies. Such measures prevent the propagation of low-quality or misleading test items and preserve the integrity of the assessment process under uncertain conditions.

## IV. DTIG COMPONENTS IN THE EDUCATIONAL SYSTEMS

Learning workflows, assessment frameworks, and automated evaluation procedures form the foundation of traditional educational assessment systems. In conventional test generation and evaluation processes, educators interact with learning content in a sequential manner by manually designing questions, reviewing syllabi, and mapping assessment items to learning objectives through predefined interfaces. These interactions involve continuous handling of instructional materials stored in structured and unstructured formats, such as textbooks, lecture notes, and digital learning resources. In educational environments with limited instructional resources or large learner populations, predefined assessment templates and scheduled content-processing routines are commonly employed to automate routine tasks such as quiz creation, grading schedules, and feedback generation. Advanced assessment systems utilize automated content pipelines and intelligent workflow engines capable of executing rule-based or ML-driven test generation procedures defined using high-level pedagogical logic and curriculum

Recent educational platforms increasingly adopt ML-driven Dynamic Test Item Generation (DTIG) mechanisms to overcome the limitations of static question banks and manual assessment design. The selection of generation and automation strategies is strongly influenced by system availability, reliability, and assessment criticality requirements. In high-stakes assessment environments, such as standardized examinations or certification testing, intelligent test generation and fault-tolerant validation mechanisms are employed to minimize assessment errors and ensure consistency, fairness, and validity. In contrast, in moderate-stakes or formative learning contexts, semi-automated DTIG approaches combined with educator oversight are often preferred to balance efficiency, pedagogical quality, and interpretability. The operational behavior of modern DTIG systems is governed by standardized learning technology and interoperability frameworks that define extensible services for content delivery, assessment generation, and learner evaluation. These frameworks support structured requests for item generation, producing corresponding test items, difficulty adjustments, or feedback outputs as system responses. Event monitoring services detect anomalies such as difficulty drift, content redundancy, or bias indicators, while action-oriented services manage automated responses including item regeneration, fallback to validated question banks, or escalation to human review. In critical situations, recovery actions may involve suspending dynamic generation and reverting to pre-validated assessments to preserve evaluation integrity. Accordingly, DTIG systems typically operate under two primary modes: an **Automated Safe Mode**, in which advanced ML-driven generation is restricted and human validation is enforced to prevent assessment risk, and an **Automated Fail-Operation Mode**, in which redundant models, alternative content sources, or conservative generation strategies are activated to maintain operational continuity. The degree of automation and autonomy within DTIG systems depends on the educational context, assessment criticality, and phase of the learning or evaluation process.

This architecture provides enhanced flexibility during both system testing and operational deployment. For instance, item quality thresholds can be dynamically adjusted, validation rules can be refined, and recovery actions can be updated to accommodate evolving curricular requirements, learner performance patterns, and assessment objectives.

## V. THE FUTURE EVOLUTION OF THE DTIG SYSTEM

Recent deployments of ML-driven Dynamic Test Item Generation (DTIG) systems in large-scale digital learning platforms have revealed several limitations in current DTIG design and development methodologies. These observations highlight opportunities for innovative approaches that complement established assessment design practices rather than replacing them entirely. This section discusses the identified deficiencies and outlines potential directions for future DTIG system evolution. Emerging educational environments increasingly demand high performance, scalability, and availability due to the rapid expansion of online learning, adaptive education, and high-stakes digital assessments. Large-scale and distributed learning systems introduce additional challenges, particularly as learner populations grow and real-time assessment requirements increase, placing greater demands on system responsiveness and reliability. Furthermore, the absence of a well-defined analytical framework to support systematic risk analysis, quality assurance, and validation activities across the DTIG lifecycle—from conceptual design to operational deployment—leads to inconsistencies across development phases and hinders the achievement of stable and robust DTIG implementations. The DTIG development toolset and validation environment therefore play a critical role in ensuring continuous support.

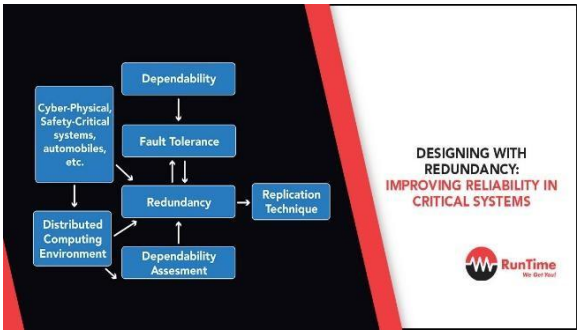



**Fig.3 concept of hardware redundancy Of DTIG**

The current approach for defining DTIG functionality typically maps one or more item generation failure modes to specific evaluation indicators, such as quality scores, confidence thresholds, or validation metrics, which trigger corrective or recovery actions when predefined limits are exceeded. However, a single failure mode may affect not only the indicator to which it is directly assigned but also other evaluation metrics used to detect different generation issues and initiate alternative recovery strategies. This situation can lead to corrective actions that fail to address the root cause of the problem during the initial DTIG stage. This limitation arises because many existing DTIG techniques rely on rigid, rule-driven generation and validation protocols. Basic validation mechanisms often treat generation symptoms independently, which may result in inconsistent conclusions, inaccurate difficulty calibration, or content misalignment. Moreover, such approaches are poorly suited to dynamic educational environments, which are time-variant, learner-dependent, and only partially observable through system monitoring signals such as learner responses or feedback patterns.

To effectively assess DTIG system health, a model-based DTIG framework is required—one that can integrate information from multiple data sources and reason about anomalous generation behavior under uncertainty, evolving learning contexts, and partial observability. Based on preliminary evaluations in real-world educational platforms, future DTIG systems should retain the

hierarchical structure described in Section II-B while incorporating more powerful and efficient detection, diagnosis, and recovery mechanisms. When selecting appropriate algorithms, three primary factors must be considered: the time required for generation fault diagnosis, the nature of the data being processed (continuous metrics, discrete events, or hybrid signals), and the computational resources required for real-time operation.

A key feature of emerging DTIG approaches is the preference for analytical redundancy over content or resource redundancy. Analytical redundancy provides a system-level view of DTIG by clearly separating the generation fault detection and identification phase from the recovery and reconfiguration phase. Traditional redundancy relies on duplicating generation logic or maintaining multiple parallel question banks to compare outputs, which increases storage and maintenance costs. In contrast, analytical redundancy detects and isolates generation faults by employing estimation techniques combined with mathematical or learned models of content quality, difficulty progression, and learner behavior. Although analytical redundancy is generally more cost-effective, it is more challenging to implement, as it must remain robust in the presence of noisy data, unpredictable learner interactions, and model uncertainty. Broadly, analytical redundancy approaches in DTIG can be classified into two categories: quantitative and qualitative model-based methods. Quantitative techniques rely on explicit models and statistical estimation to generate residuals that indicate deviations from expected generation behavior, while qualitative approaches employ artificial intelligence techniques such as pattern recognition, rule induction, and knowledge-based reasoning to identify inconsistencies and anomalies in generated test items.

The Differences between the behavior observed and the behavior anticipated by the model.

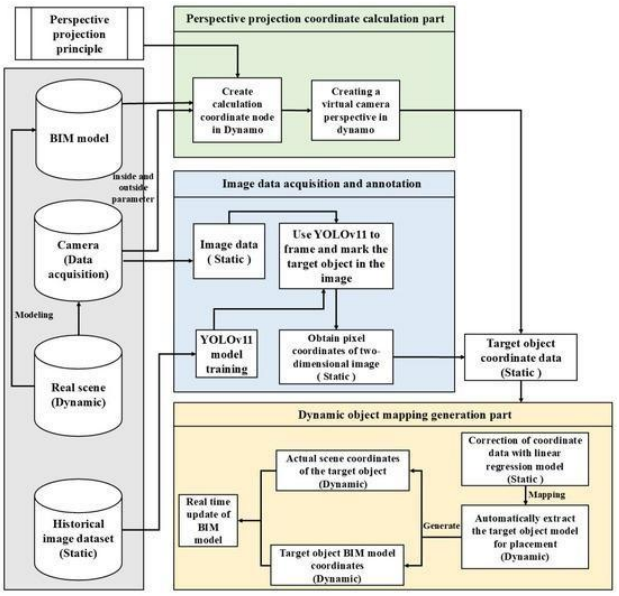



**Fig.4 OBSW architecture of autonomous system**

Fault diagnosis in DTIG systems is performed by applying appropriate decision logic after the generation and evaluation of diagnostic indicators such as quality residuals, confidence deviations, and validation scores. Autonomy is a crucial requirement for both current and future digital assessment platforms, particularly as learning systems evolve toward large-scale, adaptive, and real-time evaluation environments. Modern DTIG systems are expected to automatically interpret assessment objectives, process instructional content, and generate suitable test items even under uncertain, evolving, or partially observable learning conditions. To support this autonomy, DTIG platforms provide intuitive interfaces that allow educators or system administrators to specify high-level assessment goals, such as learning outcomes, difficulty targets, or curriculum coverage, without requiring low-level control of the generation process. To enable autonomy, planning integration, and dynamic reconfiguration, DTIG system architectures can be organized into three hierarchical levels: the decisional level, the operational level, and the functional level.

High-level assessment goals defined externally can be autonomously refined within the DTIG system into executable generation strategies and workflows for content preprocessing, item generation, and validation modules. During execution, these strategies can be dynamically adapted in response to contextual changes such as learner performance trends, updated curricula, or altered computational resource availability. DTIG functions are distributed across all three architectural levels. At the functional level, generation and validation mechanisms are closely coupled with individual content sources and processing modules, forming the foundation of the system. At the operational level, system-wide monitoring, quality control, and fault detection mechanisms analyze outputs from the functional level to identify deviations in item quality, difficulty calibration, or fairness. At the decisional level, DTIG evaluates the execution of assessment plans using aggregated information from the operational layer and detects inconsistencies between expected and observed system behavior, triggering reconfiguration or recovery actions when necessary. As optimization techniques and model-based algorithms form the foundation of modern DTIG systems, special attention must be given to their Verification and Validation (V&V). DTIG systems are often highly sensitive to variations in input data, DTIG evaluates the execution of assessment plans using aggregated information from the operational layer and detects inconsistencies between expected and observed system behavior, triggering reconfiguration or recovery actions when necessary.

- ***Runtime monitoring***:

This involves analyzing the system behavior during execution by observing generated items.

- ***Static analysis:***

Examination of DTIG software components and generation logic without execution to identify structural issues, unreachable states, or potential runtime failures.

- ***Model Checking :***

Formal verification of DTIG system behavior using executable specifications and abstract models to evaluate state evolution against defined pedagogical and quality properties.

- ***Theorem Proving:***

Application of logical reasoning and induction techniques to demonstrate that DTIG system behavior satisfies formal assessment requirements across all execution paths.

- ***Compositional Verification:***

Decomposition of DTIG system properties into properties of individual modules (e.g., preprocessing, generation, validation) to verify each component independently, thereby ensuring correctness and scalability of the overall system.

## VI. FUTURE RESEARCH

This work highlights key technical and methodological considerations for the design of Machine Learning–Driven Dynamic Test Item Generation (DTIG) systems based on insights from existing digital learning platforms and automated assessment practices. Several limitations observed in current academic and industrial approaches can be effectively addressed by integrating traditional rule-based assessment techniques with advanced AI-driven methods, including qualitative and quantitative model-based reasoning. Future educational assessment systems will increasingly demand higher levels of automation and intelligence in test item generation, with the ability to manage complex instructional content, learner diversity, and contextual uncertainty with minimal human intervention. As DTIG systems evolve, the conceptual gap between assessment design assumptions made by educators and the implicit reasoning embedded within machine learning algorithms is expected to narrow.

## VII. CONCLUSION

Machine Learning–Driven Dynamic Test Item Generation (DTIG) systems play a critical role in enhancing the efficiency, accuracy, and reliability of modern educational assessment environments. By continuously analyzing large volumes of instructional content and learner interaction data, DTIG systems are capable of generating contextually relevant, difficulty-adaptive, and pedagogically aligned test items using advanced machine learning and natural language processing techniques. These capabilities significantly reduce reliance on static question banks and manual assessment design while enabling timely and scalable evaluation.

1. **Predictive Maintenance**: By leveraging machine learning and predictive analytics, DTIG systems can anticipate potential test item quality issues—such as difficulty mismatch, ambiguity, or bias—before deployment. This enables timely corrective actions and reduces the risk of invalid or ineffective assessments.

2. **Autonomous Operations:** State-of-the-art DTIG systems support autonomous decision-making in test item generation, allowing assessments to adapt dynamically to learner performance and contextual changes without continuous human intervention.

3. **Data Integration:** State-of-the-art DTIG systems support autonomous decision-making in test item generation, allowing assessments to adapt dynamically to learner performance and contextual changes without continuous human intervention.

4. **Enhanced Mission Safety**: By detecting and mitigating generation risks at early stages, DTIG systems significantly improve the reliability, fairness, and consistency of assessments, particularly in large-scale or high-stakes educational environments.

5. **Cost Efficiency**: Early identification and resolution of test item generation issues reduce manual rework, maintenance effort.

## REFERENCES

[1] S. Kurdi, A. Leo, S. Parsia, and S. Sattler, "A Systematic Review of Automatic Question Generation for Educational Purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 121–204, 2020.

[2] M. Heilman and N. A. Smith, "Question Generation via Overgenerating Transformations and Ranking," *Language Resources and Evaluation*, vol. 44, no. 1–2, pp. 157–187, 2010.

[3] Z. A. Pardos and N. T. Heffernan, "Modeling Individualization in a Bayesian Network Implementation of Knowledge Tracing," *User Modeling and User-Adapted Interaction*, vol. 20, no. 4, pp. 365–415, 2010.

[4] X. Liu, J. Calvo, and R. A. Perez, "Automatic Item Generation Using Natural Language Processing and Machine Learning Techniques," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 742–755, 2020.

[5] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.