Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5532~5555
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.11787

Intelligent educational question–answer systems for consultation and assessment: methods, challenges, and prospects

Rakhila Turebayeva[1], Rozamgul Niyazova[2], Aktayeva Alimbubi[3], Denis Pleskachev[3], Zhanyl Sarsenbayeva[3], Musatai Esmaganbet[3]
[1]Department of Information Security, L. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Department of Artificial Intelligence Technologies, L. Gumilyov Eurasian National University, Astana, Kazakhstan
[3]Department of Information Systems and Informatics, A. Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan

Article Info

Article history:

Received Nov 4, 2025
Revised Sep 16, 2026
Accepted Oct 4, 2026

Keywords:

Adaptive testing
Cosine similarity metric
Digital learning environments
Intelligent educational systems
Item response theory
Knowledge assessment
Question difficulty modelling

ABSTRACT

Automated assessment of open-ended responses is challenging in multilingual and low-resource educational settings because annotated corpora and language-specific semantic resources are limited. Educational question–answer systems often separate psychometric modelling from natural language processing (NLP), restricting the integration of ability estimation and semantic interpretation. This study developed and evaluated an interpretable hybrid framework that combines three-parameter item response theory (IRT) adaptive testing with contextual semantic evaluation using language embeddings and cosine similarity. The framework integrates adaptive item selection, response interpretation, ability estimation, and diagnostic feedback. KazQAD, a Kazakh question-answer dataset comprising over 12,000 annotated pairs, was developed to support this study. An evaluation involving 350 undergraduates compared fixed-form computer-based testing (CBT), standalone IRT-based adaptive testing, and the hybrid framework. Among the three approaches, the hybrid framework showed the strongest correlation with expert assessment (r=0.91), the lowest mean standard error of ability estimation (SE=0.08), and the shortest average assessment duration (23 minutes vs 42 minutes for fixed-form CBT), a 45.2% reduction. These findings support integrating psychometrically calibrated estimation with semantic interpretation within a unified adaptive workflow for Kazakh-language assessment. However, this study did not directly benchmark contemporary transformer-based semantic graders, and the findings remain specific to the evaluated conditions and do not establish universal superiority.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Aktayeva Alimbubi
Department of Information Systems and Informatics, A. Myrzakhmetov Kokshetau University
Kokshetau, Kazakhstan
Email: aktaevaa2020@gmail.com

1. INTRODUCTION

The digital transformation of education has significantly expanded the role of intelligent technologies in supporting learning, assessments, and personalised instructional decision-making. The educational value of intelligent question–answer systems depends not only on automated scoring but also on their ability to identify how well an open-ended response demonstrates conceptual understanding. This requirement is especially important in multilingual educational environments, where valid knowledge may be expressed through diverse linguistic forms.

Artificial intelligence (AI), learning analytics, and adaptive educational systems have created new opportunities to deliver personalised learning experiences and improve the quality of knowledge assessment [1], [2]. In this context, intelligent educational question and answer (Q&A) systems have become an important component of modern digital learning environments because they provide automated consultations, assessments, and customised educational support.

Despite these advances, current intelligent educational assessment systems struggle to combine reliable psychometric measurements with semantic interpretations of open-ended learners' responses, particularly in multilingual, low-resource language environments. Consequently, some educational platforms provide psychometrically reliable scores but offer limited semantic interpretation of open-ended responses, whereas others provide semantic evaluations without calibrated measures of learner ability. This separation limits the reliability of automated diagnostic feedback and personalised learning support.

From the perspective of intelligent educational systems, the technical research problem is therefore the lack of a unified assessment mechanism that can jointly interpret the semantic correctness of an open-ended response and update a psychometrically calibrated estimate of learner ability. When these processes operate independently, a semantically appropriate response may not be consistently reflected in the learner model. In contrast, an ability estimate derived from structured response scoring may not capture the conceptual adequacy of the learner's explanation. This disconnect can propagate inconsistent evidence into automated diagnosis and subsequent personalised feedback. This study addresses the technical challenge of coordinating semantic response evaluation, psychometric ability estimation, adaptive item selection, and diagnostic feedback within a single assessment workflow.

Recent advances in transformer-based language models have substantially improved semantic understanding and automated response generation in educational question-answer systems. Nevertheless, most existing systems are primarily designed for information retrieval or automated grading and provide limited diagnostic feedback for open-ended learner responses because semantic and psychometric assessments are generally implemented independently.

To address some of these limitations, one promising approach to improving automated educational assessments is adaptive testing based on psychometric modelling. Item response theory (IRT) has become one of the most widely adopted frameworks for analysing learners' performances and estimating latent traits, such as ability, knowledge, and competence. Computerised adaptive testing (CAT) systems based on a three-parameter logistic model (3PL) have demonstrated strong effectiveness in personal assessment environments [3]. In this framework, the probability of a correct response is modelled as a function of the learner's abilities and item characteristics, including the difficulty, discrimination, and guessing parameters. Adaptive algorithms dynamically select test items based on the learner's estimated ability, enabling accurate measurements while reducing unnecessary testing time.

Although IRT-based adaptive testing provides a statistically reliable estimation of learner ability and efficient item selection, these models are primarily intended for structured assessment formats, such as multiple-choice (MC) questions. Consequently, they cannot adequately interpret the semantic correctness, conceptual completeness, or contextual meaning of open-ended learner responses.

While adaptive testing offers substantial psychometric advantages, educational research has highlighted the importance of metacognitive awareness in effective learning. Instruments such as the metacognitive awareness inventory (MAI) are widely used to analyse learners' self-regulated learning strategies and cognitive monitoring abilities in digital learning environments [4]. Despite these advances, most adaptive testing systems continue to rely predominantly on closed-ended question formats, such as MC items. Although these formats provide reliable psychometric measurements, they often fail to capture a deeper conceptual understanding that open-ended responses can reveal, including learners' thought processes and reasoning skills.

Consequently, intelligent educational systems are increasingly using automated methods to evaluate open-ended answers. For example, programming assessment platforms such as intelligent evaluation system using tests for teleeducation (SIETTE) extend classical test theory (CTT) and IRT principles to evaluate item reliability and validity in computational learning environments [5]. Parallel advances in natural language processing (NLP) have enabled the automated interpretation of textual responses using contextual language models. Transformer-based models such as bidirectional encoder representations from transformers (BERT) generate embeddings that capture the semantic relationships between words and sentences. In contrast, similarity metrics, such as cosine similarity, enable automated systems to evaluate conceptual correctness beyond simple keyword matching.

Building on these advances, several studies have extended psychometric and computational models to improve automated educational assessments. For example, second-order Laplace approximations have been proposed to simplify the estimation in multidimensional IRT (MIRT) models, enabling more efficient modelling of high-dimensional latent constructs. IRT has also been used to test machine learning models at the instance level, enabling the identification of data samples that are particularly challenging for predictive algorithms. Extensions such as the Γ-IRT model further extend psychometric modelling to regression

problems. In contrast, the nominal response model (NRM) provides in-depth insights into distractor behaviour in MC assessments. Recent developments in psychometrics and machine learning have broadened the analytical repertoire available for intelligent educational assessments [6]–[8].

These models clarify how different factors influence test performance and can improve assessment design and analysis. Question difficulty estimation from text (QDET) methods use linguistic and semantic features to automatically predict item difficulty and support scalable assessment content generation [9].

In smart educational technologies, accurately estimating question difficulty and providing personalised content recommendations remain important challenges. Recent research using attention-based dual multi-head co-attention mechanisms has shown potential for modelling complex relationships among questions, answer options, and explanations, improving difficulty prediction across diverse Q&A datasets [10].

Knowledge-graph-based recommender systems (KG-RS) further integrate learner profiles, educational entities, and contextual relationships to generate adaptive learning pathways, often relying on semantic similarity metrics such as cosine similarity to align educational content with learners' knowledge states [11].

Despite considerable advances in psychometric assessments and NLP, these research directions have evolved independently. Existing adaptive testing systems estimate learner ability using psychometric models, whereas NLP-based educational systems primarily evaluate the semantic similarity between textual responses. Conventional IRT-based adaptive systems provide psychometrically interpretable ability estimates, but are generally designed for structured response formats. In contrast, transformer-based semantic assessment methods support the automated analysis of open-ended responses but typically do not incorporate formal ability estimation or adaptive item selection.

The novelty of the proposed approach lies not merely in applying IRT and semantic similarity in parallel but in coordinating them within a single assessment logic that integrates item calibration, adaptive selection, semantic response interpretation, and diagnostic feedback.

Accordingly, this study addresses the following research question: to what extent does integrating IRT-based adaptive testing with contextual semantic evaluation of open-ended responses within a unified assessment workflow improve agreement with expert assessment, precision of learner ability estimation, and assessment efficiency compared with fixed-form computer-based testing (CBT) and standalone IRT-based adaptive testing?

This methodological separation creates a substantial barrier to developing intelligent educational technologies for multilingual and low-resource languages, such as Kazakh. The limited availability of annotated datasets, domain-specific linguistic resources, and pretrained language models restricts the development of advanced adaptive assessment systems. Consequently, integrated educational frameworks supported by specialised multilingual education datasets are urgently needed.

To overcome these integration challenges, this study proposes a hybrid framework for intelligent educational Q&A systems that combines psychometric adaptive testing with semantic similarity modelling in a unified digital learning environment. The proposed architecture implements CAT based on a three-parameter logistic IRT (3PL-IRT) model while simultaneously evaluating open-ended responses using contextual text embeddings and cosine similarity. This integration enables a dynamic and conceptually meaningful assessment of learners' knowledge.

The proposed framework was specifically designed to support intelligent educational technologies in low-resource language contexts, with a corpus containing more than 12,000 annotated question-answer pairs. In this study, we implemented the approach using a Kazakh-language Q&A system supported by a newly developed Kazakh question-answer dataset (KazQAD), enabling adaptive, semantically informed assessment. Unlike previous studies that treated psychometric assessment and semantic evaluation as independent processes, the proposed framework integrates both components into a unified adaptive educational architecture, thereby enabling the simultaneous estimation of learner ability and semantic interpretation of open-ended responses. The main contributions of this study can be summarised as follows:

- Hybrid psychometric-semantic assessment framework: a novel architecture integrating IRT-based adaptive testing with semantic similarity modelling for evaluating open-ended responses in intelligent educational Q&A systems.
- Multilingual adaptive assessment methodology: a methodological framework demonstrating how psychometric modelling and contextual language embedding can be combined to support adaptive knowledge diagnostics in low-resource languages.
- KazQAD development: creation of a Kazakh-language educational corpus containing more than 12,000 annotated question-answer pairs designed for intelligent multilingual assessment systems.
- Experimental validation of the hybrid assessment: an empirical evaluation with 350 university students compared the proposed hybrid framework with fixed-form CBT and standalone IRT-based adaptive testing in terms of agreement with expert assessment, precision of ability estimation, and assessment efficiency.

By integrating psychometric modelling, semantic response evaluation, adaptive item selection, and diagnostic feedback within a unified workflow, the proposed framework provides a methodological foundation for multilingual educational assessment within the evaluated setting. The framework supports personalised knowledge diagnostics and feedback while coordinating psychometric ability estimation with the semantic interpretation of open-ended learner responses.

The remainder of this paper is organised as follows: section 2 presents the theoretical background and related research. Section 3 describes the methodological framework and system architecture. Section 4 reports the experimental results and evaluates the proposed approach's performance. Section 5 discusses the implications of the findings for intelligent educational technologies and section 6 concludes the study.

2. LITERATURE REVIEW AND THEORETICAL BACKROUND

2.1. Literature review and theoretical background

Research on intelligent educational assessment integrates methods from classical psychometrics, AI, and NLP to improve the accuracy and adaptability of automated knowledge diagnostics. IRT is one of the most influential frameworks for educational measurement because it provides a probabilistic model for analyzing latent learner traits, such as knowledge, ability, and skill acquisition. Within IRT-based models, the 3PL model commonly describes the probability that a learner with a specific ability level $P_ι$ (θ) correctly answers item i, and linking item characteristics to learner performance is defined as (1):

$$P_i(\theta) = c_i + \frac{1-c_i}{1+e^{-\alpha(\theta-b_i)}} \quad (1)$$

where $\alpha_ι$ represents the item discrimination parameter, $b_ι$ denotes the item difficulty parameter, and $c_ι$ corresponds to the pseudo-guessing parameter.

The item information function (IIF), which measures the amount of information an item provides about the learner's ability level, is defined as (2):

$$I_i(\theta) = \alpha_i^2 + \frac{\left(1-P(\theta)\right)^2 P_i(\theta)}{(1-c_i)^2} \quad (2)$$

This framework enables the systematic calibration of assessment items and supports objective measurements in CBT environments. Consequently, IRT provides a theoretical foundation for CAT systems, in which items are dynamically selected based on learners' estimated ability levels [12]. This integration links item calibration, ability estimation, and adaptive item selection within a single framework.

Recent research has extended classical IRT through multidimensional, Bayesian, and testlet-based models. Studies [13], [14] examined the integration of IRT with knowledge space theory (KST), while [6] investigated second-order Laplace approximation for latent regression IRT models.

These developments enhance the scalability of large-scale intelligent assessment systems and extend classical psychometric measurements to more integrated diagnostic models that combine ability estimation with structural representations of learners' knowledge.

2.2. Semantic similarity and textual understanding

Evaluating open-ended responses in intelligent Q&A requires reliable methods for measuring the semantic similarity between learners' answers and reference solutions. Among the various approaches, cosine similarity remains one of the most widely used metrics owing to its computational efficiency and interpretability. It is defined between vector representations A and B as (3):

$$S(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} \quad (3)$$

where A and B denote the embedding vectors of the reference answer and learner's response, respectively, in (3).

In education, similarity values exceeding a predefined threshold (e.g., $S > 0.7$) are typically interpreted as semantically correct. More recently, contextual language models such as KazBERT and multilingual BERT have improved classification accuracy, reaching F1-scores close to 0.88. Together, these approaches link semantic similarity scoring with automated correctness decisions.

Several studies have proposed extensions of cosine similarity to improve robustness and applicability. For example, [15] introduced a cosine similarity ensemble (CSE) to combine multiple similarity estimators. In addition, [16] generalized cosine similarity to q-rung orthopair fuzzy sets, enabling similarity-based reasoning in decision-making frameworks such as technique for order of preference by similarity to ideal solution (TOPSIS), whereas aggregation operators such as the generalized Pythagorean fuzzy ordered weighted cosine

similarity (GPFOWCS) and the generalized Pythagorean fuzzy probabilistic ordered weighted cosine similarity (GPFPOWCS) further extended the cosine-based decision analysis [17]. Collectively, these studies indicate a shift from basic similarity scoring to a more flexible and decision-oriented similarity framework.

The scalability of similarity computation for large datasets was also addressed. For instance, a previous study [18] proposed a distributed document comparison method that uses Z-prefix indexing and Hadoop MapReduce to enable efficient similarity computation across large educational repositories.

2.3. Applications of item response theory in educational diagnostics

IRT models have been widely applied in educational diagnostics and assessment designs. One line of research has examined how item format affects diagnostic accuracy. For example, a study [19] compared MC and multiple-true-false (MTF) formats using Bayesian IRT modelling and demonstrated that item structure significantly affects measurement reliability.

Other studies have applied IRT to validate psychometric instruments and educational assessment systems. For instance, the Science Choice Preference Scale, validated using IRT [20], and the MASIA multimedia assessment system based on the Rasch model [21] demonstrated high reliability and strong discriminative capability. Intelligent learning platforms have also begun to use IRT. Based Douce *et al.* [22], the SIETTE system used both CTT and IRT models to calibrate programming tasks, enabling automatic objective performance diagnostics. The authors also examined the behavioral aspects of testing. A previous study [23] introduced testlet-based IRT models that incorporated response time to address rapid guessing behaviors. Together, these applications illustrate how IRT supports both instrument validation and adaptive diagnostic use, linking measurement validation with behavior-aware assessments.

Researchers have used multidimensional and bifactor IRT models to validate complex educational assessments. For example, one study [24] analyzed a nursing knowledge test, confirmed its factor structure, and identified biased items.

2.4. Hybrid and integrative models

The convergence of IRT, NLP, and machine learning represents an important direction for intelligent educational technologies. In this context, hybrid models aim to combine the interpretability of psychometric measurements with the representational capacity of neural architectures, linking measurement theory and computational modelling.

Models such as IRT2Vec and knowledge interaction-enhanced dynamic learner knowledge diagnosis (KIEDLKD) integrate psychometric parameters with neural memory networks to model long-term knowledge evolution and concept interactions [25]. Building on this integration, experimental tests showed that the predictions were more accurate and easier to interpret. These findings suggest that hybrid models can preserve interpretability while improving predictive performance by linking psychometric structures with neural prediction.

Other approaches combine constraint-based modelling (CBM) with psychometric models and IRT. For example, the SQL-Tutor system integrates both approaches to provide adaptive and conceptually interpretable diagnostics [26]. Similarly, Barthakur *et al.* [27] applied MIRT in large-scale MOOC environments to analyze the alignment between learning objectives and assessment items. Together, these studies show that combining psychometric and constraint-based methods can support both adaptability and interpretability while linking item-level analysis to instructional goals.

Research on automatic question generation and distractor design has benefited from psychometric integration. Kumar *et al.* [28] have demonstrated how ontologies and unstructured text can generate plausible distractors, validated using IRT-based metrics. Researchers have also used semantic similarity, lexical resources, and neural text-generation models to generate plausible incorrect answers, thus improving assessment quality [29]. Furthermore, Suh and Bolt [30] applied the NRM to analyze categorical response patterns and identify learner misconceptions. Overall, these studies linked item generation, response modelling, and diagnostic interpretation with a single assessment workflow.

2.5. Advances in psychometric modelling

2.5.1. Knowledge structure theory, cognitive, and semantic modelling

Recent research has integrated psychometric measurements with machine learning to support adaptive educational technologies. For example, one study [31] conducted a large simulation analysis of estimation methods for between-item MIRT models, comparing several estimation approaches across different sample sizes and latent factor structures.

Another important development is competence-based knowledge space theory (CbKST) [32], which focuses on diagnosing learner attributes such as skills, misconceptions, and cognitive strategies. Unlike

traditional scoring approaches, CbKST enables personalized modelling of knowledge mastery and the construction of adaptive learning pathways.

Research has also highlighted the role of metacognitive awareness in learning analytics. Consistent with this focus, Tuononen *et al.* [33] demonstrated that metacognitive awareness is closely related to students' learning profiles and self-regulated learning strategies, suggesting its potential integration with adaptive psychometric assessment models. Further studies [34], [35] extend IRT applications to validate psychometric instruments that measure conceptual understanding and statistical literacy, thereby enhancing educational assessments by ensuring that they accurately reflect students' knowledge and skills. Together, these studies suggest that psychometric assessments can better capture both the cognitive and metacognitive dimensions of learning, linking learner cognition with the measurement design.

The integration of IRT and KST has become an important approach for modelling learners' cognitive processes [32], [36]. Fundamental work on KST provides a conceptual basis for representing knowledge states and premise structures in adaptive learning environments [36]. Combining these approaches improves the modelling of learners' competencies and supports the prediction of learning trajectories.

Additional studies [19], [24] have confirmed that the cognitive structure and format of assessment items significantly influence diagnostic accuracy, indicating that well-designed assessment items can lead to more accurate evaluations of learners' competencies. Collectively, this literature positions integrated modelling as a means of improving predictive accuracy and diagnostic precision by linking learners' structures, item designs, and estimations.

Semantic similarity measurements are essential for evaluating open-ended responses. Cosine similarity remains a common baseline metric; however, several extensions have been proposed. For example, the rotary cosine similarity method presented in [37] integrates cosine similarity with intricate vector embeddings to enhance its effectiveness for completing knowledge graphs. Other studies [38], [39] have suggested cosine-based similarity measures for fuzzy sets, facilitating more adaptable models of uncertainty in textual data and bolstering decision-making frameworks, such as TOPSIS. Research by Rafiq *et al.* [40], cosine similarity measures were proposed for spherical fuzzy sets, which enabled more accurate models of uncertainty in decision-making problems. Together, these extensions move the cosine-based evaluation from a simple similarity measure to a richer uncertainty-aware model.

The proposed similarity metrics incorporate the degrees of membership, non-membership, and hesitation, enabling a more flexible representation of imprecise information in intelligent educational systems. According to Guan and Zhao [41], inner-class adjusted cosine similarity (IACS) was used to detect adversarial examples in neural networks. Ensemble methods, such as CSE and computer science and applications (CSA), improve classification robustness through adaptive thresholding and selective ensemble learning [42], [43]. To support large-scale semantic comparisons, [44] proposed a distributed framework using Z-prefix indexing and Hadoop MapReduce, enabling efficient analysis of large educational repositories. Together, these approaches extend similarity analysis from local scoring to scalable, robust educational applications.

2.5.2. Hybrid models integrating item response theory and machine learning

Recent studies have increasingly explored hybrid architectures that combine psychometric models with machine learning techniques. Idé and Dhurandhar [45] proposed an interpretable IRT-inspired predictive model for analyzing ordinal survey data using Gaussian-Hermite quadrature and expectation-maximization estimation. Lu *et al.* [46] developed an attention-based knowledge tracing model that incorporates process data and curriculum information to analyze students' learning behaviors. Their results highlight the potential of learning analytics techniques for modelling the dynamics of learner knowledge in intelligent educational systems.

Similarly, the KIEDLKD framework and related neural knowledge tracing models proposed in [47], [48] combine sequential neural architectures, cognitive modelling, and psychometric representations to capture interactions among learning tasks and track the evolution of learners' knowledge states. Several studies [49]–[51] have investigated automated approaches for programming assessment and test data generation. Reviews of automated programming evaluation methods have shown that dynamic program analysis and automatic test-case generation can significantly improve the efficiency and objectivity of grading programming assignments.

2.6. Research gap and future directions

The literature shows that IRT is a strong framework for adaptive knowledge assessment and that semantic similarity models enable the automatic interpretation of open-ended responses. However, researchers have typically developed these approaches independently. Most adaptive testing systems focus on structured responses, whereas semantic similarity models address textual evaluations without estimating psychometric ability. This separation creates a methodological gap in intelligent educational systems, particularly in multilingual contexts with limited linguistic resources.

Therefore, the methodological challenge is not simply to integrate IRT and NLP modules into the same system. Rather, it requires establishing coherent assessment logic, in which adaptive item selection, semantic response interpretation, learner ability estimation, and diagnostic feedback are coordinated in a transparent, interpretable workflow.

The challenge becomes more pronounced for low-resource languages such as Kazakh, where annotated datasets and pretrained language models remain scarce. To address this gap, this study proposes an intelligent Q&A framework that integrates the following:

- Psychometric adaptive testing based on IRT.
- Semantic evaluation of open-ended responses using similarity metrics.
- Interactive consultation mechanisms providing pedagogically meaningful feedback.

By combining psychometric interpretation with semantic analysis and adaptive consultation mechanisms, the proposed framework aims to advance the development of transparent, scalable, and learner-centered intelligent educational systems. Based on this gap, the present study proposes a hybrid IRT-semantic modelling framework that integrates psychometric adaptive testing with semantic response evaluation.

Conventional IRT-based adaptive systems provide psychometrically interpretable ability estimates, but are typically designed for structured response formats. In contrast, transformer-based semantic assessment methods support automated analysis of open-ended responses but generally do not incorporate formal ability estimation or adaptive item selection.

The methodological gap lies in the lack of a unified psychometric-semantic workflow that supports adaptive testing, interpretable ability estimation, and semantically informed diagnostic feedback for open-ended learners' responses. The study is organized around the following progression: the educational context and diagnostic problem; the respective capabilities and limitations of psychometric and semantic approaches; the methodological gap arising from their separation; the proposed hybrid framework; and its significance for multilingual, low-resource educational environments.

3. MATERIALS AND METHOD

3.1. Research framework

This section presents the methodological framework for developing the proposed intelligent educational Q&A system. The methodology integrates IRT-based psychometric modelling, vector-based semantic similarity evaluation, and an adaptive consultation mechanism supported by the KazQAD.

This integrated approach aims to ensure accurate knowledge diagnostics while maintaining semantic interpretability and pedagogical relevance in multilingual digital learning environments. The methodological architecture combined the following three complementary components:

- Psychometric assessment layer based on IRT for estimating learner ability.
- Semantic evaluation layer using embedding-based similarity metrics.
- An adaptive consultation layer providing personalized feedback and learning recommendations.

Together, these components form a hybrid architecture that supports both formative and summative assessments in digital education systems. Specifically, the Q&A system employs a multilayer design that integrates psychometric modelling, semantic analysis, and adaptive consulting. The framework consists of three interconnected analytical levels.

- Psychometric level (IRT): this layer quantitatively estimates the learner's latent ability parameter using adaptive testing based on a 3PL model.
- Semantic level (NLP): this module evaluates the semantic similarity between the learner's answer and the reference solution using vector representations derived from contextual language models.
- Adaptive-consultative level (KazQAD): this level manages question sequencing, knowledge tracing, and the generation of personalized pedagogical feedback using the KazQAD.

Integrating these layers enables a unified learning environment in which assessment, semantic interpretation, and feedback generation operate within a single adaptive workflow. The architecture integrates the following three major components:

- IRT-based adaptive testing engine for estimating learner ability.
- A semantic evaluation module based on BERT embeddings and cosine similarity was used to analyze the open-ended responses.
- The KazQAD dataset provides annotated question-answer pairs and psychometric parameters, as shown in Figure 1.

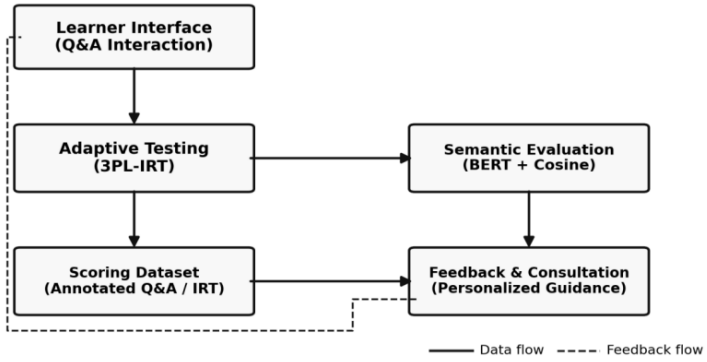


Figure 1. System architecture of the proposed hybrid IRT-semantic educational Q&A system

Figure 1 presents the architecture of the proposed hybrid educational question-answer system. The learner interface supports interactive Q&A activities, whereas the adaptive testing engine applies the 3PL IRT model to select assessment items and estimate learners' abilities. The semantic evaluation module analyses open-ended responses using BERT-based embeddings and cosine similarity. The annotated scoring dataset integrates question-answer records with IRT parameters. In contrast, the feedback and consultation engine generates personalized recommendations based on the learner's updated ability estimate and semantic correctness of the responses.

The annotated scoring dataset stores question–answer records together with calibrated IRT parameters. In contrast, the feedback and consultation engine generates personalized recommendations based on the learner's updated ability estimate and the semantic correctness of the submitted responses.

3.2. Psychometric model based on the three-parameter item response theory

The core of the proposed assessment model is the 3PL-IRT model. The probability that a learner with a given ability level θ correctly answers item i is defined as (4):

$$P_i(\theta) = c_i + \frac{1-c_i}{1+e^{-a_i(\theta-b_i)}} \quad (4)$$

where α_ι is the discrimination parameter; b_ι is the difficulty parameter; and c_ι is the pseudo-guessing parameter.

The model is characterized by three parameters: the discrimination parameter, which measures an item's sensitivity to variations in learner abilities; the difficulty parameter, which indicates the ability level at which the probability of a correct response equals 0.5; and the pseudo-guessing parameter, which represents the probability of correctly answering an item by guessing.

Parameter estimation is performed using maximum likelihood estimation (MLE) or Bayesian approximation through iterative calibration. For a learner with an observed response vector $U = (u_1,..., u_n)$, the likelihood function is expressed as (5):

$$L(\theta \mid U) = \prod_{i=1}^{n} [P_i(\theta)]^{u_i} [1 - P_i(\theta)]^{1-u_i}, \quad (5)$$

where $u_i \in \{0,1\}$ denotes the scored response to item i, $P_i(\theta)$ is the probability of a correct response under the 3PL-IRT model, and θ denotes learner ability.

Maximizing $(L(\theta|U))$ yields the estimated learner ability parameter θ. The IIF is defined as (6):

$$I_i(\theta) = a_i^2 \frac{(1-P_i(\theta))^2 P_i(\theta)}{(1-c_i)^2} \quad (6)$$

The next item in an adaptive sequence is selected by maximizing the information function (7):

$$i^* = arg \max_i I_i(\theta) \quad (7)$$

This procedure ensures that each selected question contributes maximally to the precision of ability estimation.

3.3. Adaptive testing algorithm

The system implements a CAT strategy that dynamically selects questions based on the learner's performance. The adaptive testing process is as follows:

- Initialization: set the initial ability level $\theta_\iota=0$.
- Item selection: choose the item ι that maximizes $I_i(\theta)$.
- Response collection: record learner response (u_j).
- Termination condition: stop if the standard error (SE) $(\theta)\leq 0.01$ the threshold is reached or the maximum number of items is exceeded.
- Ability update: after each scored response, the learner's ability distribution was updated using the Bayesian posterior based on the 3PL-IRT likelihood:

$$p(\theta|u_{1:t})=\frac{p(\theta)\prod_{j=1}^{t}[P_j(\theta)]^{u_j}[1-P_j(\theta)]^{1-u_j}}{\int_{-\infty}^{+\infty}p(\theta)\prod_{j=1}^{t}[P_j(\theta)]^{u_j}[1-P_j(\theta)]^{1-u_j}\,d\theta} \quad (8)$$

where $u_{1:t}=(u_1,\ldots,u_t)$ denotes the sequence of scored learner responses up to iteration t, $u_j\in\{0,1\}$, $P_j(\theta)$ is the probability of a correct response to item j under the 3PL-IRT model, and $p(\theta)$ is the prior distribution of learner ability. The denominator normalizes the posterior distribution so that it integrates to one. The updated posterior distribution was then used to obtain the current ability estimate $\hat{\theta}_t$, for example by the expected a posteriori estimate: $\hat{\theta}_t=\int_{-\infty}^{+\infty}\theta\ p(\theta|u_{1:t})\,d\theta$.

To retain the dichotomous response structure required by the 3PL model, the semantic score was converted into an IRT response indicator. A response was assigned ($u_j = 1$) when ($S(R_j,A_j) \geq 0.70$); otherwise, it was assign toed ($u_j = 0$). Responses with ($0.40 \leq S(R_j, A_j) < 0.70$) were retained as partially correct for diagnostic feedback, while the continuous similarity score was also preserved for qualitative interpretation.

- Feedback generation: compute a diagnostic report and personalized learning recommendations.

Empirical studies indicate that this adaptive approach can reduce the test length by 30-40% while maintaining measurement precision, which is particularly important in remote learning environments. The system utilizes calibrated IRT parameters with discrimination values $\alpha \in [0.5, 2.5]$ to ensure adequate diagnostic sensitivity.

3.4. Semantic evaluation of open-ended responses

To evaluate open-ended responses, the system applies cosine similarity between the contextual embedding of the reference answer and the contextual embedding of the learner response as (9):

$$S(R_i, A_i) = \frac{\mathbf{e}(R_i)^{\top}\mathbf{e}(A_i)}{\|\mathbf{e}(R_i)\|_2\|\mathbf{e}(A_i)\|_2}, \quad (9)$$

where R_i denotes the reference answer for item i, A_i denotes the learner's response, and $e(\cdot)$ denotes the contextual embedding generated by KazBERT or multilingual BERT. The similarity score $S(R_i, A_i)$ ranges from -1 to 1, with higher values indicating greater semantic correspondence between the learner response and the reference answer. The similarity score $S(R_i, A_i)$ is interpreted as follows:

- $S(R_i, A_i)<0.4$: semantically inconsistent (misunderstanding or irrelevant answer).
- $0.40\leq S(R_i, A_i)<0.70$: partially correct (incomplete understanding).
- $S(R_i, A_i)\geq 0.70$: semantically correct responses.

To improve the robustness of the semantic evaluation, a weighted cosine similarity approach was applied using term-level weights, such as term frequency–inverse document frequency (TF-IDF) weights. The weighted similarity between the reference answer R_i, and learner response A_i is defined as (10):

$$S_w(R_i, A_i)=\frac{\sum_{j=1}^{m} w_j, r_{ij}, a_{ij}}{\sqrt{\sum_{j=1}^{m} w_j, r_{ij}^2}, \sqrt{\sum_{j=1}^{m} w_j, a_{ij}^2}}, \quad (10)$$

where r_{ij} denotes the value of the j-th semantic or lexical feature in the embedding representation of the reference answer, R_i, a_{ij} denotes the corresponding feature value in the learner response A_i, w_j represents the weight of the j-th feature based on its discriminative value in the subject domain, and m is the number of analysed features.

This formulation ensures consistency with (9) because both the standard and weighted semantic similarity measures compare the learner's response with the reference answer rather than with the question

text. The weighted formulation improves the evaluation of multi-word responses and domain-specific terms by assigning greater importance to conceptually informative aspects.

For example, for an item on big data, the reference answer may define big data as data whose volume, velocity, or complexity exceeds the capacity of conventional processing tools. A learner response such as "data volumes that exceed processing capacity" may receive a high KazBERT-based similarity score (e.g., $S(R_i,A_i)=0.92$) because it preserves the core conceptual meaning of the reference answer. In contrast, a more general response, such as "a significant amount of information," may receive a lower similarity score (e.g., $S(R_i,A_i)=0.78$) because it captures the notion of data volume but omits the processing-capacity criterion. This difference reflects variation in conceptual precision rather than simple keyword overlap.

Most conventional computer-based and adaptive testing systems are optimized for selected-response items. Therefore, they provide limited evidence regarding the semantic adequacy, conceptual completeness, and contextual relevance of open-ended learner responses in general.

3.5. Architecture of the Kazakh question–answer dataset and hybrid assessment workflow

KazQAD provides an annotated educational corpus containing more than 12,000 question–answer pairs enriched with psychometric parameters, cognitive-level labels, and semantic metadata. The corpus covers several academic domains, including computer science, mathematics, physics, cybersecurity, and humanities. KazQAD directly underpins the workflow of the proposed hybrid framework in a Kazakh-language, low-resource educational setting by supporting psychometric calibration, semantic response evaluation, adaptive assessment, diagnostic feedback, and experimental validation. Each KazQAD record was annotated in three complementary dimensions:

- Psychometric metadata include the IRT parameters a, b, and c, representing item discrimination, difficulty, and pseudo-guessing characteristics, respectively.
- Each item was classified according to Bloom's taxonomy.
- Semantic metadata included keywords, concept tags, reference answers, and contextual text representations that support the evaluation of learners' responses.

3.5.1. Reference-answer construction and linguistic variation

For reference-answer construction and linguistic variation, each assessment item in KazQAD includes a primary reference answer and semantic metadata, including keywords and concept tags. The reference answer represents the item's expected conceptual content and serves as the semantic comparison target. Learner responses are not evaluated by exact lexical matching; instead, contextual embeddings generated by KazBERT or multilingual BERT are compared using cosine similarity. Therefore, paraphrases, inflectional variants, and lexically different formulations can still obtain high similarity scores when they preserve the core conceptual meaning.

However, the available dataset documentation does not provide sufficient detail to reconstruct a formal multi-expert reference-answer validation protocol. Accordingly, this study does not make claims about inter-rater agreement or a documented multi-expert reference-answer validation process. In addition, the current implementation uses a primary reference answer rather than an exhaustive set of alternative valid answers. This limitation is particularly relevant for higher-order open-ended questions, where several conceptually correct formulations may exist. Future versions of KazQAD should therefore include multiple independently validated reference answers and paraphrastic variants for each item.

The present study did not involve training or fine-tuning a new language model. Therefore, KazQAD was not divided into conventional training, validation, or test subsets for supervised encoder training. Instead, the semantic diagnostic component was evaluated using stratified five-fold cross-validation of the annotated response records. Each fold preserved the distribution of the three response correctness categories: correct, partially correct, and incorrect. This procedure enabled the evaluation of the semantic classification performance within the hybrid workflow while maintaining the representativeness of the response categories. The KazQAD architecture comprises four functional layers.

- The data layer stores annotated question–answer pairs and associated metadata.
- The semantic layer contains contextual vector representations generated using KazBERT and multilingual BERT models.
- The assessment layer implements IRT-based psychometric modelling and semantic similarity evaluation.
- Finally, the feedback layer generates personalized learning recommendations based on learner ability estimates and evidence from correct semantic responses, linking assessment results to feedback.

Table 1 presents representative KazQAD records and illustrates the integrated annotation structure of the corpus. Each record combines a question–answer pair with psychometric parameters, Bloom's taxonomy labels, and semantic information. This representation enables the coordinated implementation of adaptive item

selection, semantic interpretation of open-ended responses, learner ability estimation, and personalized feedback within the proposed hybrid educational Q&A system, making the workflow connections explicit.

Table 1. Example records from the kazakh question–answer dataset (KazQAD)

ID	Question (KZ)	Answer (KZ)	Bloom level	a	b	c
153	«*Ақпараттық қауіпсіздік дегеніміз не*? - What is information security?"	«*Бұл ақпараттың тұтастығын, құпиялылығын және қолжетімділігін сақтау* - Maintaining confidentiality, integrity, and availability of information"	Understand	1.85	-0.23	0.18
221	"What is encryption?"	"The process of encoding information to prevent unauthorised access."	Apply	2.12	0.45	0.12

KazQAD's layered structure supports the integration of psychometric calibration, semantic evaluation, and adaptive feedback mechanisms. Accordingly, the corpus serves as an educational knowledge base and is the primary empirical resource for evaluating the proposed hybrid IRT-semantic assessment workflow, with each layer contributing sequentially to the overall process.

3.6. Cognitive classification using Bloom's taxonomy

KazQAD supports classifying assessment items according to the six cognitive levels defined in Bloom's taxonomy.

- Recall of factual knowledge and definitions.
- Understanding: explanation of concepts and ideas.
- Application of knowledge to practical problems.
- Analysis: identification of relationships and structural components.
- Evaluation: critical assessment and judgement.
- Create: generation of new ideas, solutions, or conceptual structures.

The psychometric parameters estimated using the IRT demonstrated a consistent relationship with these cognitive levels. In particular, the difficulty parameter b progressively increased from lower cognitive levels (remember and understand) to higher levels (evaluate and create). Simultaneously, the discrimination parameter tended to reach its maximum value for tasks involving analytical reasoning and synthesis, reflecting its stronger diagnostic capability in distinguishing learning ability levels. Such an alignment between psychometric calibration and cognitive taxonomy supports the validity of the KazQAD as a dataset for adaptive educational assessment.

3.7. Hybrid assessment workflow

The integrated assessment process combines psychometric modelling and semantic response evaluation in a unified adaptive workflow. The assessment procedure begins by initializing the learner's ability parameters. Subsequently, the system selects the most informative items by maximizing the IRT information function. After each response was collected, the learner's ability estimate was updated, semantic similarity was evaluated, and feedback was generated accordingly.

We generated contextual representations using KazBERT and multilingual BERT as embedding-based feature extractors. The semantic similarity between the reference answer and the learner's response was computed using cosine similarity. The resulting semantic scores were mapped to response correctness categories using predefined thresholds: semantically incorrect, partially correct, and semantically correct. The system then uses the categorical result as response evidence for updating ability and retains the continuous similarity score for diagnostic feedback.

After the learner submits a response, the system evaluates its semantic correctness using contextual text embedding and cosine similarity. The system then incorporates the resulting similarity score into the psychometric model to update the learner's ability estimate. This iterative process was continued until the termination criterion was satisfied.

Figure 2 illustrates the workflow of the proposed hybrid IRT semantic assessment framework. The process integrates adaptive item selection based on IRT with the semantic evaluation of open-ended responses using contextual language embeddings. The resulting hybrid workflow combines psychometric measurement, semantic interpretation, and personalized feedback within a unified adaptive assessment process for digital learning environments.

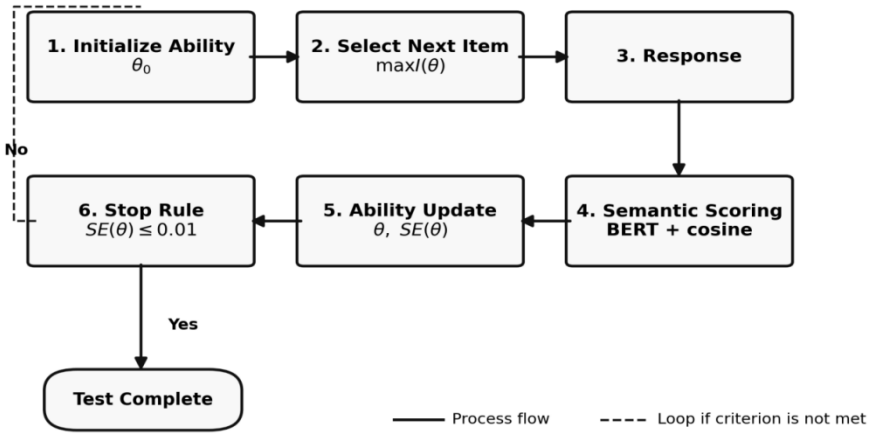


Figure 2. Workflow of the hybrid IRT-semantic assessment cycle

The workflow consisted of the following stages:

- Initialization of learner ability parameter.
- Adaptive item selection using the IRT information function.
- Learner response submission.
- Semantic evaluation of response using cosine similarity.
- Ability to update using Bayesian estimation.
- Termination of assessment.

This iterative cycle continued until a stable estimate of the learner's ability was obtained. Through this integrated workflow, the hybrid architecture combines the following features:

- Psychometric measurement precision.
- Semantic interpretation of open-ended responses.
- Adaptive learning system (ALS).

To illustrate the methodological positioning of the proposed approach, Table 2 compares representative educational assessment methods with the hybrid IRT-semantic framework across psychometric modelling, semantic evaluation, adaptive item selection, support for open-ended responses, multilingual applicability, and interpretability.

Conventional IRT-based adaptive systems provide psychometrically interpretable ability estimates but are typically designed for structured response formats. In contrast, transformer-based semantic assessment methods support automated analysis of open-ended responses but generally do not incorporate formal ability estimation or adaptive item selection. The proposed framework addresses these complementary limitations by integrating 3PL-IRT-based adaptive testing, embedded semantic evaluation, and personalized diagnostic feedback. This contrast motivates the experimental evaluation of agreement with expert assessment, precision of ability estimation, semantic classification performance, and assessment efficiency presented in the following section.

Table 2. Comparative analysis of the proposed hybrid IRT-semantic framework and representative educational assessment approaches

Approach	Psychometric modelling	Semantic evaluation	Adaptive testing	Open-ended response assessment	Personalized feedback
Conventional CBT	✓	✗	✗	✗	✗
IRT-based adaptive testing	✓	✗	✓	✗	Limited
Transformer-based semantic assessment (e.g., BERT/SBERT)	✗	✓	✗	✓	Limited
Proposed hybrid IRT-semantic framework	✓	✓	✓	✓	✓

As shown in Table 2, the proposed framework provides a unified architecture that combines psychometric measurement, semantic interpretation, adaptive testing, and personalized feedback. Unlike approaches based solely on psychometric testing or semantic response classification, this integration enables a more comprehensive assessment of learners' knowledge and highlights the contribution of the framework.

Despite considerable advances in psychometric assessments and NLP, these research directions have generally evolved independently. This study addresses this separation by coordinating adaptive item selection, semantic response evaluation, ability estimation, and personalized feedback within a single multilingual assessment workflow.

The comparison shows that the proposed approach combines psychometric modelling, semantic evaluation, and adaptive consultation mechanisms within a unified architecture. Consequently, the proposed approach provides an interpretable framework for integrating psychometric modelling, semantic evaluation, and adaptive consultation mechanisms within the evaluated multilingual assessment setting. The evaluation of the proposed structure under the present experimental conditions was performed using the KazQAD system and a sample of university students.

4. RESULTS AND EXPERIMENTAL EVALUATION

4.1. Experimental design

This section presents an experimental evaluation of the proposed hybrid intelligent Q&A system. The evaluation was conducted from psychometric, semantic, and educational perspectives to assess the integrated psychometric-semantic architecture's performance in measurement accuracy, semantic validity, and adaptive assessment efficiency. Using KazQAD and involving 350 undergraduate students from three universities in Kazakhstan, this study covered informatics, information security, and educational IT programs. The participants were assigned to three experimental groups for comparison, as shown in Table 3.

Measurement accuracy: comparative results in Group C (IRT+NLP) showed that the correlation coefficient between the estimated ability level and expert evaluation reached r=0.91, compared with r=0.82 in Group B and r=0.75 in Group A. The mean SE of ability estimation decreased from 0.19 to 0.08, indicating greater precision than in the other groups.

Table 3. Comparative psychometric performance and assessment efficiency across assessment conditions

No	Assessment condition	Correlation with expert evaluation (r)	Mean SE	Average duration (min)	Time reduction vs. CBT (%)
1	Fixed-form CBT	0.75	0.19	42	0.0
2	IRT-based adaptive testing	0.82	0.13	29	31.0
3	Hybrid IRT-semantic framework	0.91	0.08	23	45.2

From an instructional perspective, these results indicate that learners can be assessed more efficiently if they receive richer diagnostic information. In addition, the hybrid assessment condition was associated with an approximately 45% reduction in the average assessment time relative to fixed-form CBT, indicating greater assessment efficiency under the evaluated conditions.

The experimental design enabled a comparative analysis of traditional testing approaches and the proposed hybrid intelligent assessment system. The evaluation used KazQAD, which contains 12,000 annotated question-answer pairs across multiple academic domains.

The experimental configuration was designed to ensure reproducibility and consistency across all evaluation stages. Specifically, semantic representations of learner responses were generated using contextual embeddings from the KazBERT and multilingual BERT models, and semantic similarity was computed using cosine similarity.

The comparative evaluation included fixed-form CBT and standalone IRT-based adaptive testing as direct psychometric baselines. The present study did not include a direct empirical comparison with SBERT, XLM-R, or fine-tuned transformer-based semantic grading models. Accordingly, interpret the reported results as evidence of the added value of integrating semantic evaluation into an IRT-based adaptive assessment workflow, rather than as a state-of-the-art benchmark against all contemporary NLP-based assessment systems.

Therefore, the corpus was not partitioned for supervised language model training. However, the diagnostic classification performance was estimated using stratified five-fold cross-validation of the annotated evaluation records.

Psychometric ability estimation was performed using the 3PL-IRT model with Bayesian updating after each learner response. Adaptive item selection followed the maximum item information criterion, and the assessment ended when the SE of the ability estimate met the stopping criterion or when the predefined maximum number of assessment items was reached. Thus, the hybrid evaluation combines psychometric ability estimation with semantic response analysis within a unified adaptive assessment workflow.

4.2. Measurement accuracy and comparative performance

As shown in Table 3, the proposed hybrid IRT-semantic framework showed higher agreement with expert assessment and lower ability-estimation error than conventional CBT and standalone IRT-based adaptive testing. Combining psychometric modelling and semantic response evaluation was associated with the highest observed agreement with expert assessment and the lowest observed SE of ability estimation among the compared assessment conditions.

To ensure a comprehensive evaluation of the proposed system, several quantitative metrics were employed to assess the classification performance and psychometric ability estimation. The evaluation included the accuracy, precision, recall, F1-score, mean absolute error (MAE), and SE of the estimated learner ability parameter. Accuracy, precision, recall, and F1-score were used for semantic response classification, whereas MAE and SE were used for the psychometric model's precision. Correlation analysis between estimated learner ability and expert assessment scores was used to assess agreement with expert evaluation.

Because only aggregate group-level results were available for this analysis, the individual-level data and within-group dispersion statistics needed for reliable retrospective calculation of confidence intervals and standardized effect-size estimates were not available. Accordingly, the confidence intervals and effect sizes were not reconstructed from the aggregate values reported in Table 3. Future evaluations will retain participant-level results to enable uncertainty estimation and effect-size reporting.

Table 4 summarizes the confirmed performance indicators of the hybrid IRT-semantic framework. The confirmed performance indicators include agreement with expert evaluations, SE of ability estimation, MAE, multiclass semantic classification accuracy, average assessment duration, and time reduction compared with fixed-form CBT. Together, these measures evaluate the psychometric precision and semantic classification performance of the proposed framework.

For the Kazakh-language subcorpus, the semantic classification component achieved an average F1-score of 0.83 and an accuracy of 0.86. These findings indicate that combining adaptive psychometric modelling with contextual semantic analysis can support the evaluation of open-ended learner responses in the investigated Kazakh-language educational setting.

Table 4. Confirmed performance indicators of the hybrid IRT-semantic framework

Performance indicator	Value	Evaluation scope
Correlation with expert evaluation (r)	0.91	Overall hybrid assessment framework
Mean SE of ability estimation	0.08	3PL-IRT ability estimation
MAE of ability estimation	0.12	Psychometric estimation component
Multiclass semantic classification accuracy	0.88	Fully annotated evaluation set; stratified five-fold cross-validation
Average assessment duration	23 min	Hybrid assessment condition
Time reduction relative to fixed-form CBT	45.2%	Derived from mean assessment duration

4.3. Semantic validity and linguistic robustness

The semantic validity of the proposed framework was evaluated by assessing the ability of the semantic evaluation module to classify open-ended learners' responses into three categories of response correctness: correct, partially correct, and incorrect. Unlike keyword-based grading, the proposed approach evaluates the semantic correspondence between the reference answer and learner response using contextual embeddings generated by KazBERT and multilingual BERT.

For each learner response, semantic similarity was computed between the contextual embedding of the reference answer and that of the learner response, as defined in (9). The resulting similarity score was mapped to one of the three response-correctness categories using predefined thresholds: incorrect, partially correct, and correct. This categorization enabled the system to evaluate not only fully correct answers but also incomplete or partially adequate responses, which are common in open-ended educational assessments.

The classification performance was evaluated using multiclass accuracy, which is defined as the proportion of learner responses assigned to the correct response-correctness category:

$$\text{Accuracy}=\frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i=y_i)\frac{\sum_{k=1}^{3} n_{kk}}{N} \quad (11)$$

where N is the total number of evaluated responses, y_i denotes the reference class of response i, $\hat{y}_i$ denotes the predicted class, $(I(\cdot))$ is the indicator function, and n_{kk} represents the number of responses correctly assigned to the category k. These three categories correspond semantically to correct, partially correct, and incorrect responses.

Across the fully annotated evaluation set, the semantic classification component achieved an overall multiclass accuracy of 0.88 under stratified five-fold cross-validation. This value represents the semantic

evaluation module's aggregate performance across all annotated response records included in the evaluation protocol.

A separate analysis was conducted on the Kazakh-only response subset to assess linguistic robustness under morphologically complex and low-resource language conditions. In this subset, the semantic classification component achieved 0.86 accuracy and an average F1-score of 0.83. The slightly lower accuracy can be explained by the linguistic characteristics of Kazakh-language learner responses, including inflectional variation, brief answers, metaphorical expressions, and dialectal forms, on which current contextual embedding models are insufficiently trained.

Table 5 summarizes the semantic classification model's category-level performance across the full annotated evaluation set. The F1-scores ranged from 0.84 to 0.89. The highest F1-score was obtained for correct responses (0.89), whereas partially correct and incorrect responses achieved F1-scores of 0.84 and 0.85, respectively. These results indicate that the semantic evaluation module can distinguish between fully correct, partially correct, and incorrect open-ended learner responses with acceptable consistency.

Table 5. Performance of the semantic classification model across response-correctness categories

Response category	Average F1-score
Correct responses	0.89
Partially correct responses	0.84
Incorrect responses	0.85

The KazBERT-based semantic component demonstrated that contextual embeddings can support the evaluation of morphologically complex Kazakh-language learner responses in the present experimental setting. However, these results should not be interpreted as direct evidence of superiority over established English-language datasets or contemporary transformer-based semantic graders because no direct benchmarking against SQuAD, RACE, SBERT, XLM-R, or fine-tuned semantic assessment models was conducted. Rather, the findings support the feasibility of applying contextual semantic evaluation to open-ended learner responses in Kazakh language educational assessments.

4.4. Cognitive validation of the Kazakh question-answer datasets and efficiency of the hybrid adaptive system

To evaluate the psychometric validity of the KazQAD, the relationship between Bloom's cognitive levels and the IRT difficulty parameter (b) was examined. Pearson's correlation coefficient was calculated as a supplementary measure of the linear association between the numerically coded Bloom's taxonomy level and item difficulty, as (12):

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \quad (12)$$

where x_i represents the cognitive level of an assessment item, y_i denotes the corresponding IRT difficulty parameter b, and n is the number of analyzed items.

The Spearman rank correlation coefficient was calculated to account for possible non-linear monotonic relationships as (13):

$$\rho_s=\frac{\sum_{i=1}^{n}(R_{x_i}-\overline{R_x})(R_{y_i}-\overline{R_y})}{\sqrt{\sum_{i=1}^{n}(R_{x_i}-\overline{R_x})^2}\sqrt{\sum_{i=1}^{n}(R_{y_i}-\overline{R_y})^2}} \quad (13)$$

where R_{x_i} and R_{y_i} denote the ranks of Bloom's taxonomy level and IRT difficulty parameter, respectively.

This formulation accommodates tied ranks, which are expected because multiple assessment items may belong to the same Bloom's cognitive level. The difference between the ranks of the paired observations was examined, providing the basis for the following correlation analysis.

The analysis revealed a strong positive correlation between Bloom's cognitive level and the IRT difficulty parameter, where is the Spearman correlation coefficient, and p is the significance level. These findings indicate that items requiring higher cognitive processes tended to correspond to larger values of the difficulty parameter.

Tasks associated with higher levels of Bloom's taxonomy, such as analyze, evaluate, and create, require greater learner ability than tasks involving lower-level cognitive processes, such as remember or understand.

An experimental analysis was conducted on the KazQAD dataset, which contains more than 12,000 annotated question-answer pairs and covers six cognitive levels of Bloom's taxonomy. The empirical evaluation involved 350 undergraduate students whose responses were used to estimate item parameters and learner ability values within the 3PL IRT model.

Table 6 summarizes the relationship between the cognitive levels defined by Bloom's taxonomy and the corresponding IRT parameters for the KazQAD assessment items. The results indicate that the discrimination parameter generally increases with cognitive complexity, suggesting that higher-level tasks differentiate learners with varying ability levels more effectively.

Table 6. Statistical relationship between Bloom's levels and IRT parameters

No.	Bloom's level	Mean discrimination (a)	Mean difficulty (b)	Avg. success rate (%)
1	Remember	0.95	-1.20	92
2	Understand	1.35	-0.60	81
3	Apply	1.85	0.10	74
4	Analyze	2.10	0.55	68
5	Evaluate	2.45	0.80	61
6	Create	2.80	1.25	55

The difficulty parameter also showed a progressive increase across Bloom's levels, thus supporting the psychometric consistency of the dataset. These findings provide evidence that item parameters align with the intended cognitive structure of assessment tasks.

This indicates that higher-order assessment items provide stronger diagnostic power than lower-order items in distinguishing learners with different ability levels. Simultaneously, the average success rate decreased across Bloom's levels, reflecting the increased difficulty of the tasks that require deeper cognitive processing.

Figure 3 illustrates the relationship between Bloom's taxonomy levels and the IRT difficulty parameter, showing a monotonic increase across cognitive levels. This figure demonstrates a clear positive trend between cognitive complexity and item difficulty, providing empirical support for the psychometric validity of the KazQAD. In addition to validating the cognitive structure of the datasets, an experimental evaluation was conducted to assess the efficiency of the proposed hybrid adaptive assessment.

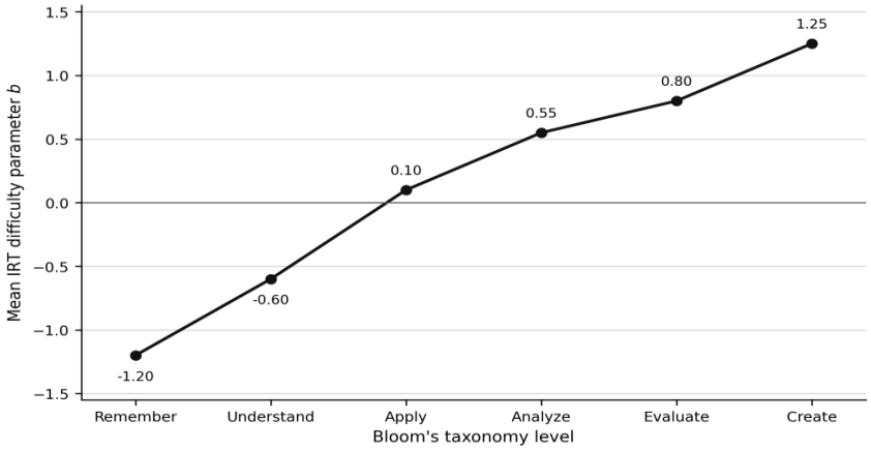


Figure 3. Relationship between Bloom's taxonomy levels and the IRT difficulty parameter b

The hybrid assessment workflow combines IRT-based adaptive item selection and semantic response evaluation in two complementary ways. First, adaptive item selection reduces unnecessary item administration by selecting questions that provide the greatest psychometric information at the learner's current estimated ability level. Second, semantic evaluation provides immediate interpretation of open-ended learner responses and can support faster scoring. In the experimental evaluation, the hybrid framework was associated with an approximately 45% reduction in average assessment duration compared with fixed-form CBT.

The efficiency of the adaptive assessment process is determined by the cumulative information provided by the administered items. At iteration (t), the test information and the conditional SE of ability estimation are defined as (14):

$$I_T(\hat{\theta}_t)=\sum_{j\in A_t} I_j(\hat{\theta}_t), \quad SE(\hat{\theta}_t)=\frac{1}{\sqrt{I_T(\hat{\theta}_t)}}, \tag{14}$$

where A_t denotes the set of items administered up to iteration t, $I_j(\hat{\theta}_t)$ is the information provided by item j at the current ability estimate, and $SE(\hat{\theta}_t)$ represents the conditional SE of the current ability estimate. The assessment was terminated when the predefined SE threshold was reached or when the maximum number of assessment items had been administered.

Adaptive item selection prioritizes the items expected to provide the largest amount of information in the learner's current ability estimate. Consequently, the system can reduce unnecessary item administration while maintaining the target precision of ability estimation. Automated semantic evaluation complements this process by providing immediate interpretation of written responses, thereby supporting faster scoring, efficient assessment completion, and timely feedback to learners.

The psychometric properties of the adaptive testing process are illustrated in Figure 4, which presents the item characteristic curve (ICC) and the corresponding IIF for the 3PL-IRT model. ICC shows the probability of a correct response across different learner ability levels, whereas the IIF indicates the range of ability values for which the item provides the greatest measurement precision.

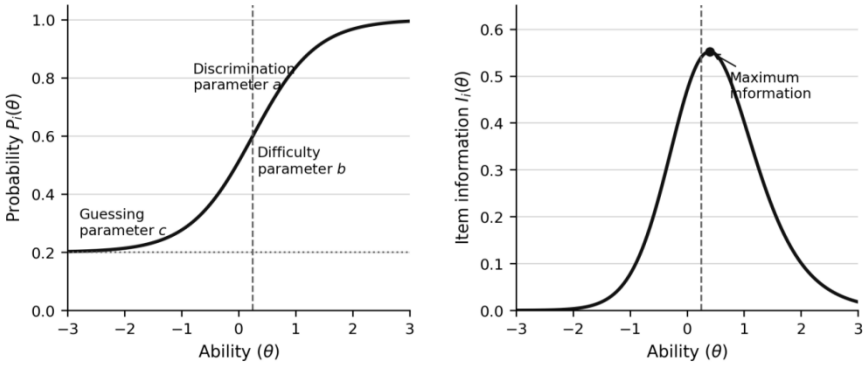


Figure 4. ICC and IIF in the 3PL IRT model

Figure 4 illustrates the psychometric behavior of the 3PL model and the role of item information in adaptive item selection. The results indicate that the items were selected to maximize measurement precision at the learner's current estimated ability level, which supports the effectiveness of the proposed adaptive assessment procedure. ICC illustrates the probability of a correct response as a function of the learner's ability parameter, whereas IIF indicates the range of ability levels in which the assessment item provides the maximum diagnostic information.

Figure 5 compares the agreement with expert evaluation across fixed-form CBT, standalone IRT-based adaptive testing, and the proposed hybrid IRT semantic framework. The results show that the hybrid framework achieved the highest correlation with expert assessment, indicating a stronger alignment between automated ability estimation and expert judgement.

Figure 5 illustrates the degree of agreement with the expert evaluation across the assessment conditions. The hybrid IRT-semantic framework achieved the highest correlation with expert evaluation (r=0.91) compared with standalone IRT-based adaptive testing (r=0.82) and fixed-form CBT (r=0.75). The hybrid IRT-semantic assessment condition showed a stronger alignment with expert judgment than the two psychometric baseline conditions under the evaluated conditions.

The observed performance advantage can be attributed to the complementary roles of the components of the framework. The IRT module selects items that are maximally informative for the learner's current estimated ability level, thereby improving measurement efficiency and reducing unnecessary item administration.

The semantic evaluation module determines whether the content of an open-ended learner response aligns with the reference concept, providing diagnostic evidence unavailable in conventional selected-response testing. Together, these components support a more informative learner profile and enable the generation of targeted feedback for learners.

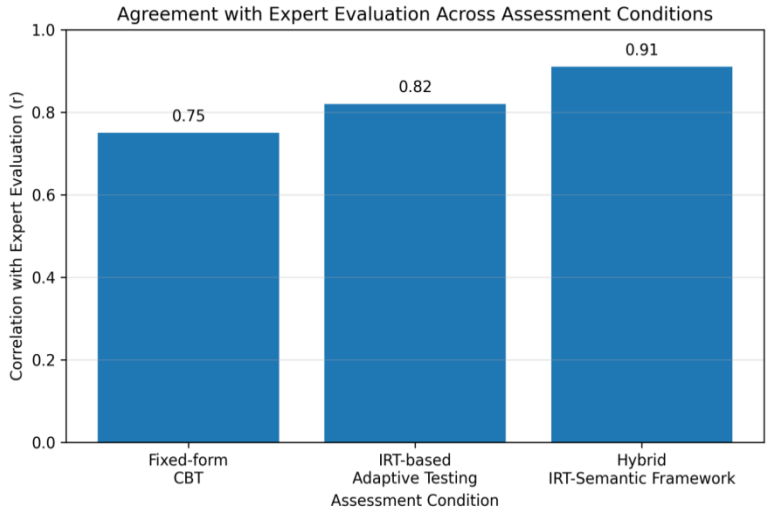


Figure 5. Agreement with expert evaluation across assessment conditions

4.5. Interpretability of the psychometric model

An important advantage of the proposed hybrid framework is its interpretability. Unlike many black-box neural models, the IRT component provides a mathematically transparent representation of learner performance through explicit item parameters and an estimated ability level. In this framework, each assessment result is accompanied by a diagnostic visualization that illustrates the probability of a success curve and the corresponding (IIF), as shown in Figure 6.

Figure 6 illustrates the relationship between learner ability, probability of a correct response, and item information in the 3PL IRT model. The probability curve represents the likelihood that a learner with an ability level correctly answers item *i*. A lower asymptote reflects a pseudo-guessing parameter, whereas the probability of a correct response increases as the learner's ability increases.

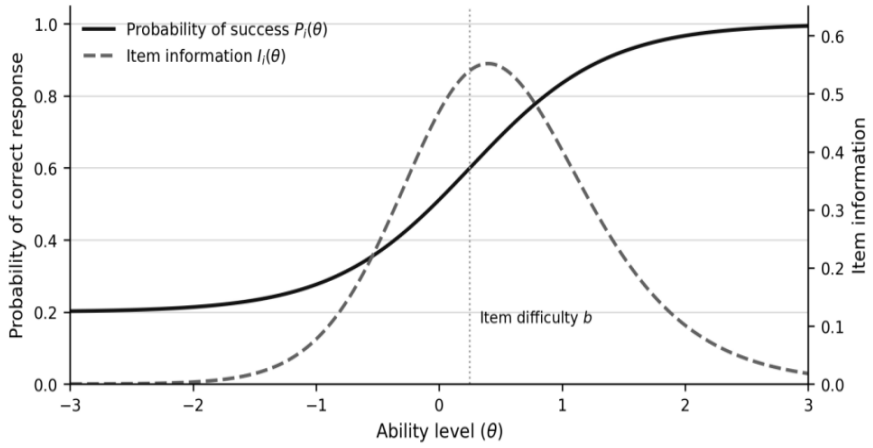


Figure 6. Probability of success curve and IIF in the 3PL IRT model

The IIF $I_i(\theta)$ indicates the range of ability values for which an item provides the highest measurement precision. In the 3PL model, the maximum information is typically obtained near and may occur slightly above the item difficulty parameter b. Therefore, items calibrated at different difficulty levels contribute to accurate ability estimations across a broad range of learner competencies. These psychometric representations enhance transparency in the assessment process and support the informed adjustment of item parameters in adaptive learning environments.

4.6. Limitations of the current model

This study presents a methodological framework with experimental validation and does not claim to describe a production-ready institution-wide system. Despite the overall effectiveness of the proposed framework, several limitations should be acknowledged, and each point leads to a specific next step:

- The system demonstrated lower accuracy when evaluating very short responses, particularly single-word answers, because it provided limited semantic information for contextual similarity analysis. Future studies should address this limitation.
- Approximately 7% of the misclassified responses involved metaphorical expressions or dialectal variations of the Kazakhs that were insufficiently represented in the current KazBERT training corpus. Future studies should expand the representation of these forms.
- Semantic ambiguity remains a challenge for tasks at the "Create" level of Bloom's taxonomy, where multiple conceptually valid responses may be expressed in different linguistic forms. Future work should refine the handling of such ambiguities.

The empirical evaluation was limited to the selected Kazakh-language corpus, academic domains, and participating student population; therefore, future work should validate the findings across additional disciplines, learner groups, and multilingual educational contexts.

Although the framework uses contextual representations generated by KazBERT and multilingual BERT, the present study did not include direct empirical benchmarking against recent transformer-based semantic assessment models, such as SBERT, XLM-R, or task-specific fine-tuned semantic graders. In addition, no ablation analysis was conducted to quantify the individual contributions of psychometric adaptation, semantic evaluation, and personalised feedback to the framework's overall performance.

These limitations highlight several priorities for future research. Further work should expand KazQAD to include additional academic domains, dialectal variants, and multiple reference answers; evaluate the framework using independent learner samples; compare it with contemporary transformer-based semantic assessment models under a common experimental protocol; and conduct ablation and uncertainty analyses to assess the contributions of the individual components of the framework.

5. DISCUSSION

The experimental results indicate that combining psychometric modelling with semantic evaluation can enhance the reliability and efficiency of automated knowledge diagnostics in an educational setting. Unlike conventional CAT systems, which rely predominantly on structured response formats, the proposed framework evaluates open-ended learner responses while retaining explicit item parameters and learner ability estimates. This combination is particularly relevant in multilingual educational environments, where learners may express conceptually valid knowledge through diverse linguistic forms.

The observed performance advantage stems from the complementary roles of the framework components. The IRT module selects items that are maximally informative for the learner’s current estimated ability level, thereby improving measurement efficiency and reducing unnecessary item administration. The semantic evaluation module determines whether an open-ended learner response aligns with the reference answer, providing diagnostic evidence that conventional selected-response tests cannot capture. Together, these components provide more informative learner profiles and enable targeted feedback.

In the present experimental evaluation, the hybrid framework showed greater agreement with expert assessments and a lower ability-estimation error than conventional CBT and standalone IRT-based adaptive testing. The hybrid framework achieved the highest agreement with expert assessment ($r=0.91$), followed by standalone IRT-based adaptive testing ($r=0.82$), and fixed-form CBT ($r=0.75$). It also reduced the average assessment duration from 42 to 23 min, representing a 45.2% reduction compared to conventional fixed-form CBT. These findings suggest that integrating adaptive psychometric assessments with semantic interpretations can support more precise and efficient evaluations of learners’ knowledge.

The observed improvement in agreement with expert evaluation likely reflects the complementary roles of the two components rather than the effect of either component alone. IRT-based adaptive testing increases measurement efficiency by selecting items that provide greater information at the learner’s current estimated ability level, thereby reducing the administration of less informative items. Semantic evaluation contributes a different source of evidence by assessing the conceptual correspondence of open-ended responses rather than relying solely on selected-response correctness. Within the proposed workflow, semantically evaluated responses are incorporated into the psychometric ability-updating process, allowing evidence from open-ended responses to contribute to the learner model. This integration plausibly explains the stronger agreement with expert evaluation and shorter assessment duration observed under the hybrid condition. However, because the present study did not include an ablation analysis isolating each component's

contribution, the current results cannot quantify the relative contribution of adaptive item selection and semantic evaluation independently.

These findings are consistent with previous educational assessment studies that demonstrated the use of IRT in computer-adaptive and adaptive learning environments [3], [13]. They also complement recent NLP-based educational assessment research showing that contextual language models and sentence-embedding approaches can support the automated evaluation of open-ended student responses and achieve meaningful agreement with human graders [52]. Unlike approaches that examine psychometric adaptation and semantic response evaluation separately, this study evaluates their integration within a unified workflow in which semantic evidence from open-ended responses updates psychometric ability.

This study does not claim that the proposed framework is superior to all transformer-based educational assessment models. The empirical comparison focused on fixed-form CBT and standalone IRT-based adaptive testing because they serve as the principal psychometric baselines for evaluating the added value of semantic response interpretation. Direct benchmarking against SBERT, multilingual BERT, XLM-R, and fine-tuned semantic grading models is necessary to assess the proposed framework's performance relative to contemporary NLP-based educational assessment approaches.

From a practical perspective, the framework can support formative assessment by helping teachers distinguish between incorrect, partially correct, and semantically adequate responses. It may also reduce the effort required to repeatedly evaluate open-ended tasks and provide learners with timely feedback regarding conceptual gaps. The KazQAD offers an initial resource for educational administrators and policymakers seeking to develop Kazakh-language digital assessment tools and extend inclusive educational services to low-resource linguistic communities.

The scope of the present evaluation constrains the generalisability of these findings. The empirical study involved 350 undergraduate students from three universities in Kazakhstan and focused on a limited set of academic domains represented in the current evaluation. Consequently, the observed results should not be assumed to generalise directly to other languages, educational levels, academic disciplines, institutions, or learner populations. Validation using broader multilingual corpora, additional academic domains, and more diverse participant populations is therefore required.

This study's broader contribution supports the development of intelligent educational technologies for linguistically underrepresented populations. High-resource languages benefit from extensive corpora, pre-trained language models, and established digital assessment platforms, whereas low-resource languages often lack comparable educational resources. By combining an annotated Kazakh-language corpus with an interpretable psychometric-semantic framework, this study demonstrated a feasible methodological pathway for multilingual adaptive assessment.

Future research should expand the KazQAD across additional academic domains, incorporate multiple reference answers and more advanced contextual semantic models, evaluate speech-based question-answering interfaces, and investigate gamified adaptive learning mechanisms. It should also assess cross-linguistic transfer to other underrepresented languages, benchmark more systematically against transformer-based assessment models, and conduct an ablation analysis of individual framework components to evaluate the proposed approach's scalability and generalisability.

6. CONCLUSION

This study's main contribution is the development and empirical evaluation of a unified multilingual assessment framework that combines psychometric modelling, adaptive item selection, contextual semantic evaluation of open-ended responses, and personalised diagnostic feedback in a single assessment workflow. Rather than treating psychometric measurement and semantic response interpretation as separate processes, the proposed architecture coordinates these components so that semantic evidence from learner responses can contribute to ability estimation and subsequent adaptive assessment decisions. This integration is the principal methodological contribution of this study and provides a framework for combining measurement-oriented and language-aware assessment mechanisms in the Kazakh language educational context.

The empirical evaluation involving 350 undergraduate students showed that, under the evaluated conditions, the hybrid IRT-semantic framework achieved a correlation of r=0.91 with expert assessments, a mean SE of ability estimation of 0.08, and an average assessment duration of 23 min, compared with 42 min for fixed-form CBT. The semantic classification component achieved an overall multiclass accuracy of 0.88 under stratified five-fold cross-validation. These findings provide empirical evidence that integrating psychometric ability estimation, adaptive item selection, and contextual semantic response evaluation is feasible and effective within the evaluated assessment workflow.

These findings should be interpreted in light of the current coverage of the KazQAD corpus, semantic ambiguity in open-ended responses, reduced reliability for very brief or dialectal answers, and the absence of direct benchmarking against recent transformer-based semantic assessment models. Accordingly, the present

results should not be interpreted as evidence of universal superiority over contemporary semantic assessment approaches. In addition, the evaluation was limited to the selected institutional and linguistic contexts; therefore, external validation across additional learner populations, disciplines, and languages is required to determine the framework's broader applicability.

Future research should expand the KazQAD across academic domains, incorporate multiple reference answers and advanced contextual semantic models, release reproducible implementation materials, conduct ablation and uncertainty analyses, and compare the framework with contemporary transformer-based assessment methods. Speech-based Q&A interfaces and cross-linguistic transfers to other low-resource languages should also be investigated. These extensions will enable a more rigorous evaluation of the generalisability and comparative performance of the proposed hybrid architecture.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rakhila Turebayeva		✓	✓		✓	✓		✓	✓		✓		✓	✓
Rozamgul Niyazova	✓	✓			✓	✓		✓	✓	✓		✓		
Aktayeva Alimbubi	✓	✓	✓	✓			✓		✓	✓	✓		✓	
Denis Pleskachev			✓			✓	✓	✓		✓	✓			
Zhanyl Sarsenbayeva			✓		✓		✓			✓		✓		✓
Musatai Esmaganbet				✓	✓		✓	✓	✓		✓			✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

INFORMED CONSENT STATEMENT

Informed consent was obtained from all subjects involved in the study.

ETHICAL APPROVAL

The study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee at the Department of Artificial Intelligence (AI) Technologies, L.N.Gumilyov Eurasian National University (protocol code 2025/488 and date of approval 13 October 2025).

DATA AVAILABILITY

The original contributions of this study are included in this article and the data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] B. Hymer, D. Michel, and L. Todd, "Dynamic consultation: Towards process and challenge," *Educational Psychology in Practice*, vol. 18, no. 1, pp. 47–62, Mar. 2002, doi: 10.1080/02667360120122813.
[2] A. Farea and F. Emmert-Streib, "Understanding question-answering systems: Evolution, applications, trends, and challenges," *Engineering Applications of Artificial Intelligence*, vol. 156, Sep. 2025, doi: 10.1016/j.engappai.2025.110997.
[3] A. Huda, F. Firdaus, D. Irfan, Y. Hendriyani, A. Almasri, and M. Sukmawati, "Optimizing educational assessment: The practicality

of computer adaptive testing (CAT) with an item response theory (IRT) approach," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 1, pp. 473–480, Mar. 2024, doi: 10.62527/joiv.8.1.2217.
[4] G. M. Harrison and L. M. Vallin, "Evaluating the metacognitive awareness inventory using empirical factor-structure evidence," *Metacognition and Learning*, vol. 13, no. 1, pp. 15–38, Apr. 2018, doi: 10.1007/s11409-017-9176-z.
[5] R. Conejo, B. Barros, and M. F. Bertoa, "Automated assessment of complex programming tasks using SIETTE," *IEEE Transactions on Learning Technologies*, vol. 12, no. 4, pp. 470–484, Oct. 2019, doi: 10.1109/TLT.2018.2876249.
[6] B. Andersson and T. Xin, "Estimation of latent regression item response theory models using a second-order Laplace approximation," *Journal of Educational and Behavioral Statistics*, vol. 46, no. 2, pp. 244–265, Apr. 2021, doi: 10.3102/1076998620945199.
[7] J. V. C. Moraes, J. T. S. Reinaldo, M. Ferreira-Junior, T. S. Filho, and R. B. C. Prudêncio, "Evaluating regression algorithms at the instance level using item response theory," *Knowledge-Based Systems*, vol. 240, Mar. 2022, doi: 10.1016/j.knosys.2021.108076.
[8] T. I. Smith and N. Bendjilali, "Motivations for using the item response theory nominal response model to rank responses to multiple-choice items," *Physical Review Physics Education Research*, vol. 18, no. 1, Apr. 2022, doi: 10.1103/PhysRevPhysEducRes.18.010133.
[9] L. Benedetto *et al.*, "A survey on recent approaches to question difficulty estimation from text," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–37, Sep. 2023, doi: 10.1145/3556538.
[10] H.-J. Song, S.-H. Yoon, and S.-B. Park, "Question difficulty estimation based on an attention model for question answering," *Applied Sciences*, vol. 11, no. 24, Dec. 2021, doi: 10.3390/app112412023.
[11] C. Troussas, A. Krouska, P. Tselenti, D. K. Kardaras, and S. Barbounaki, "Enhancing personalized educational content recommendation through cosine similarity-based knowledge graphs and contextual signals," *Information*, vol. 14, no. 9, Sep. 2023, doi: 10.3390/info14090505.
[12] D. Rizopoulos, "ltm: An R package for latent variable modeling and item response theory analyses," *Journal of Statistical Software*, vol. 17, no. 5, 2006, doi: 10.18637/jss.v017.i05.
[13] P. J. Muñoz-Merino, R. G. Novillo, and C. D. Kloos, "Assessment of skills and adaptive learning for parametric exercises combining knowledge spaces and item response theory," *Applied Soft Computing*, vol. 68, pp. 110–124, Jul. 2018, doi: 10.1016/j.asoc.2018.03.045.
[14] S. Noventa, A. Spoto, J. Heller, and A. Kelava, "On a generalization of local independence in item response theory based on knowledge space theory," *Psychometrika*, vol. 84, no. 2, pp. 395–421, Jun. 2019, doi: 10.1007/s11336-018-9645-6.
[15] P. Xia, L. Zhang, and F. Li, "Learning similarity with cosine similarity ensemble," *Information Sciences*, vol. 307, pp. 39–52, Jun. 2015, doi: 10.1016/j.ins.2015.02.024.
[16] D. Liu, X. Chen, and D. Peng, "Some cosine similarity measures and distance measures between q -rung orthopair fuzzy sets," *International Journal of Intelligent Systems*, vol. 34, no. 7, pp. 1572–1587, Jul. 2019, doi: 10.1002/int.22108.
[17] R. Verma and A. Mittal, "Multiple attribute group decision-making based on novel probabilistic ordered weighted cosine similarity operators with Pythagorean fuzzy information," *Granular Computing*, vol. 8, no. 1, pp. 111–129, Jan. 2023, doi: 10.1007/s41066-022-00318-1.
[18] M. Alewiwi, C. Orencik, and E. Savaş, "Efficient top-k similarity document search utilizing distributed file systems and cosine similarity," *Cluster Computing*, vol. 19, no. 1, pp. 109–126, Mar. 2016, doi: 10.1007/s10586-015-0506-0.
[19] C. E. Brassil and B. A. Couch, "Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison," *International Journal of STEM Education*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40594-019-0169-0.
[20] L. Sha, C. Schunn, and M. Bathgate, "Measuring choice to participate in optional science learning experiences during early adolescence," *Journal of Research in Science Teaching*, vol. 52, no. 5, pp. 686–709, May 2015, doi: 10.1002/tea.21210.
[21] C.-Y. Kuo, H.-K. Wu, T.-H. Jen, and Y.-S. Hsu, "Development and validation of a multimedia-based assessment of scientific inquiry abilities," *International Journal of Science Education*, vol. 37, no. 14, pp. 2326–2357, Sep. 2015, doi: 10.1080/09500693.2015.1078521.
[22] C. Douce, D. Livingstone, and J. Orwell, "Automatic test-based assessment of programming," *Journal on Educational Resources in Computing*, vol. 5, no. 3, Sep. 2005, doi: 10.1145/1163405.1163409.
[23] K.-Y. Jin, C.-L. Hsu, M. M. Chiu, and P.-H. Chen, "Modeling rapid guessing Behaviors in computer-based testlet items," *Applied Psychological Measurement*, vol. 47, no. 1, pp. 19–33, Jan. 2023, doi: 10.1177/01466216221125177.
[24] Y. Cai, "The value of using test response data for content validity: An application of the bifactor-MIRT to a nursing knowledge test," *Nurse Education Today*, vol. 35, no. 12, pp. 1181–1185, Dec. 2015, doi: 10.1016/j.nedt.2015.05.014.
[25] W. Gan, Y. Sun, and Y. Sun, "Knowledge interaction enhanced sequential modeling for interpretable learner knowledge diagnosis in intelligent tutoring systems," *Neurocomputing*, vol. 488, pp. 36–53, Jun. 2022, doi: 10.1016/j.neucom.2022.02.080.
[26] J. Gálvez, E. Guzmán, R. Conejo, A. Mitrovic, and M. Mathews, "Data calibration for statistical-based assessment in constraint-based tutors," *Knowledge-Based Systems*, vol. 97, pp. 11–23, Apr. 2016, doi: 10.1016/j.knosys.2016.01.024.
[27] A. Barthakur, S. Joksimovic, V. Kovanovic, M. Richey, and A. Pardo, "Aligning objectives with assessment in online courses: Integrating learning analytics and measurement theory," *Computers & Education*, vol. 190, Dec. 2022, doi: 10.1016/j.compedu.2022.104603.
[28] A. P. Kumar, A. Nayak, M. S. K., S. Goyal, and Chaitanya, "A novel approach to generate distractors for multiple choice questions," *Expert Systems with Applications*, vol. 225, Sep. 2023, doi: 10.1016/j.eswa.2023.120022.
[29] Y. Susanti, T. Tokunaga, H. Nishikawa, and H. Obari, "Automatic distractor generation for multiple-choice English vocabulary questions," *Research and Practice in Technology Enhanced Learning*, vol. 13, no. 1, Dec. 2018, doi: 10.1186/s41039-018-0082-z.
[30] Y. Suh and D. M. Bolt, "Nested logit models for multiple-choice item response data," *Psychometrika*, vol. 75, no. 3, pp. 454–473, Sep. 2010, doi: 10.1007/s11336-010-9163-7.
[31] M. Garnier-Villarreal, E. C. Merkle, and B. E. Magnus, "Between-item multidimensional IRT: How far can the estimation methods go?," *Psych*, vol. 3, no. 3, pp. 404–421, Aug. 2021, doi: 10.3390/psych3030029.
[32] P. Anselmi, J. Heller, L. Stefanutti, and E. Robusto, "Educational and clinical assessment within competence-based knowledge structure theory," *Testing, Psychometrics, Methodology in Applied Psychology*, vol. 32, no. 1, pp. 1–18, 2025, doi: 10.69978/Tpmap.V32.I1.1.
[33] T. Tuononen, H. Hyytinen, M. Räisänen, T. Hailikari, and A. Parpala, "Metacognitive awareness in relation to university students' learning profiles," *Metacognition and Learning*, vol. 18, no. 1, pp. 37–54, Apr. 2023, doi: 10.1007/s11409-022-09314-x.
[34] A. Rachmatullah *et al.*, "Development and validation of the middle grades computer science concept inventory (MG-CSCI) assessment," *EURASIA Journal of Mathematics, Science and Technology Education*, vol. 16, no. 5, Feb. 2020, doi: 10.29333/ejmste/116600.
[35] A. Sabbag, A. Zieffler, and C. Ng, "Can we distinguish statistical literacy and statistical reasoning?," *Statistics Education Research*

Journal, vol. 24, no. 1, pp. 1–19, Jan. 2025, doi: 10.52041/serj.v24i1.587.
[36] J.-P. Doignon and J.-C. Falmagne, "Spaces for the assessment of knowledge," *International Journal of Man-Machine Studies*, vol. 23, no. 2, pp. 175–196, Aug. 1985, doi: 10.1016/S0020-7373(85)80031-6.
[37] L. Wang, J. Luo, S. Deng, and X. Guo, "RoCS: Knowledge graph embedding based on joint cosine similarity," *Electronics*, vol. 13, no. 1, Dec. 2023, doi: 10.3390/electronics13010147.
[38] M. Rahim, S. S. Abosuliman, R. Alroobaea, K. Shah, and T. Abdeljawad, "Cosine similarity and distance measures for quasirung orthopair fuzzy sets: Applications in investment decision-making," *Heliyon*, vol. 10, no. 11, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32107.
[39] G. Wei, "Some cosine similarity measures for picture fuzzy sets and their applications to strategic decision making," *Informatica*, vol. 28, no. 3, pp. 547–564, Jan. 2017, doi: 10.15388/Informatica.2017.144.
[40] M. Rafiq, S. Ashraf, S. Abdullah, T. Mahmood, and S. Muhammad, "The cosine similarity measures of spherical fuzzy sets and their applications in decision making," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 6, pp. 6059–6073, Jun. 2019, doi: 10.3233/JIFS-181922.
[41] D. Guan and W. Zhao, "Adversarial detection based on inner-class adjusted cosine similarity," *Applied Sciences*, vol. 12, no. 19, Sep. 2022, doi: 10.3390/app12199406.
[42] F. F. Altıntaş, "A new method for measuring the performance of alternatives: Cosine similarty approach," *Politeknik Dergisi*, vol. 28, no. 6, pp. 1691–1705, Dec. 2025, doi: 10.2339/politeknik.1627566.
[43] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv Preprint*, Feb. 2014, doi: 10.48550/arXiv.1306.6709.
[44] T. Elsayed, J. Lin, and D. W. Oard, "Pairwise document similarity in large collections with MapReduce," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08*, 2008, pp. 265–268, doi: 10.3115/1557690.1557767.
[45] T. Idé and A. Dhurandhar, "Supervised item response models for informative prediction," *Knowledge and Information Systems*, vol. 51, no. 1, pp. 235–257, Apr. 2017, doi: 10.1007/s10115-016-0976-2.
[46] Y. Lu, Y. Cheng, and L. Tong, "Advanced knowledge tracing: Incorporating process data and curricula information via an attention-based framework," *Journal of Educational Data Mining*, vol. 16, no. 2, pp. 58–84, 2024.
[47] Y. Liang, T. Peng, Y. Pu, and W. Wu, "HELP-DKT: an interpretable cognitive model of how students learn programming based on deep knowledge tracing," *Scientific Reports*, vol. 12, no. 1, Mar. 2022, doi: 10.1038/s41598-022-07956-0.
[48] X. Zhou, Z. Zhang, X. Xie, and J. Zhang, "Deep learning based knowledge tracing in intelligent tutoring systems," *Scientific Reports*, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-07422-7.
[49] R. Romli, N. Mahzan, M. Mahmod, and M. Omar, "Test data generation approaches for structural testing and automatic programming assessment: A systematic literature review," *Advanced Science Letters*, vol. 23, no. 5, pp. 3984–3989, May 2017, doi: 10.1166/asl.2017.8294.
[50] S. Combéfis, G. Beresnevičius, and V. Dagienė, "Automated code assessment in programming education: A review," *Computers*, vol. 10, no. 1, 2022.
[51] R. Romli, S. Sulaiman, and K. Z. Zamli, "Automatic programming assessment and test data generation a review on its approaches," in *2010 International Symposium on Information Technology*, Jun. 2010, pp. 1186–1192, doi: 10.1109/ITSIM.2010.5561488.
[52] J. Pecuchova, Ľ. Benko, and M. Drlik, "Automated grading of open-ended questions in higher education using GenAI models," *International Journal of Artificial Intelligence in Education*, vol. 35, no. 6, pp. 3813–3846, Dec. 2025, doi: 10.1007/s40593-025-00517-2.

BIOGRAPHIES OF AUTHORS



Rakhila Turebayeva candidate of physical and mathematical sciences and associate professor, presented at L.N. Gumilyov Eurasian National University, Department of Information Security. The leader's scientific interests include mathematical modelling of systems, optimization methods, and mathematical modelling of discrete systems. She can be contacted at email: turebayeva_rd@enu.kz.



Rozamgul Niyazova graduated from the Faculty of Information Technology of Kh. Dosmukhamedov Atyrau State University offers a degree in physics and computer science. In 2010, she defended her thesis for the degree of Candidate of Technical Sciences, specialty 05.13.11 – "Mathematical and software of computers, complexes and computer networks" – on the topic "Methods for improving the performance of a distributed corporate information system". In 2011, she presented at L.N. Gumilyov Eurasian National University, Department of Theoretical Computer Science, as an associate professor of artificial intelligence technologies. Her research interests include artificial intelligence, machine learning, and intelligent learning systems. She can be contacted at email: rs.niyazova@gmail.com.



Aktayeva Alimbubi obtained bachelor's and master's degrees in Applied Mathematics from S. M. Kirov Kazakh State University, as well as a Ph.D. in Computer Science from T. Ryskulov Kazakh Economic University (Kazakhstan). She is currently a professor in the Department of Information Systems and Informatics at Abay Myrzakhmetov University in Kokshetau. She has published several research articles in prestigious conference proceedings and journals. Her research interests include quantum computing, cybersecurity and intelligent information systems. She is actively involved in research and peer review, with a particular focus on applying cutting-edge technologies to address real-world security challenges in intelligent environments and critical sectors. In addition, she supervises postgraduate students and acts as a reviewer and keynote speaker at international events. She can be contacted at email: aktaevaa2020@gmail.com or aakhtaewa@gmail.com.



Denis Pleskachev is a senior lecturer at A. Myrzakhmetov Kokshetau University, Kazakhstan. He holds an M.Sc. in Computer Science from A. Myrzakhmetov Kokshetau University. His research interests include text mining, machine learning, natural language processing, and online social networks. He can be contacted at email: denispleskachev@mail.ru.



Zhanyl Sarsenbayeva is a senior lecturer at A. Myrzakhmetov Kokshetau University, Kazakhstan. She holds an M.Sc. in Computer Science from A. Myrzakhmetov Kokshetau University. Her research interests include text mining, machine learning and online social networking. Through her scholarly work, she actively conducts research to advance information systems and technology management. She can be contacted at email: zhanyl.sarsenbayeva@mail.ru.



Musatai Esmaganbet currently serves as an associate professor at A. Myrzakhmetov Kokshetau University, Kazakhstan. His primary research focuses on networking and wireless networking, particularly in detecting hate speech, cybercrime, and cyberbullying behaviour within online social networks. He can be contacted at email: esmaganbet_m@mail.ru.