Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979−995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

Combining Generative AI and Scheduling Algorithms for Personalized Learning Powered by TELISIK

Artamananda[1,*], Eogenie Lakilaki[2], Syakillah Nachwa[3], Muhammad Gilang Ramadhan[4], Amaliah Sobli[5], Annisa Fatihah Salsabila[6], Bagus Ramadhan[2]

[1]PT Vanz Inovatif Teknologi, Jakarta
Ciputra International, Jalan Lingkar Luar Barat No. 101, Lantai 17 Unit 2 & 5, Rawa Buaya, Kecamatan Cengkareng, Kota Jakarta Barat, DKI Jakarta, Indonesia
[2]National Library of The Republic of Indonesia, Jakarta
Jl. Medan Merdeka Sel. No.11, Gambir, Kecamatan Gambir, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta, Indonesia
[3]Faculty of Computer Science, Information System, Universitas Sriwijaya, Ogan Ilir
Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra, Indonesia
[4]PT Bukalapak.com Tbk, Jakarta
Treasury Tower Lantai 65, District 8 SCBD, Jl. Jendral Sudirman Kav. 52–53, Jakarta Selatan, DKI Jakarta, Indonesia
[5]Regional Representative Council of the Republic of Indonesia, Palembang
Jl. Kapten A. Rivai No.1, Lorok Pakjo, Kec. Ilir Bar. I, Kota Palembang, Sumatera Selatan, Indonesia
[6]SIT Robbani, Ogan Ilir
Jl. Sarjana No.Blok A, Timbangan, Kecamatan Indralaya Utara, Kabupaten Ogan Ilir, Sumatera Selatan, Indonesia
Email: [1,*]artamananda@gmail.com, [2]eogenie@perpusnas.go.id, [3]chwakillah@gmail.com,
[4]gilang.ramadhan@multirealmgames.com, [6]annisafatihahsalsabila@gmail.com, [7]bagus_ramadhan@perpusnas.go.id,
Correspondence Author Email: artamananda@gmail.com
Submitted: **19/06/2026**; Accepted: **14/07/2026**; Published: **14/07/2026**

Abstract−The increasing competitiveness of the Computer Based Written Examination for National Selection Based on Test (UTBK-SNBT) necessitates adaptive and sustainable learning support. This study proposes a web based intelligent learning platform that integrates Artificial Intelligence (AI) to facilitate examination preparation through Automatic Question Generation (AQG), an AI Tutor, automated solution generation, and a scheduler driven question generation mechanism. The platform adopts a client server architecture and employs a Large Language Model (LLM) to generate examination questions tailored to the characteristics of each UTBK-SNBT subtest. The system was evaluated from two perspectives: the learning performance of 30 students across seven UTBK-SNBT subtests and the quality of 1.663 AI generated questions using a Jaccard Similarity based deduplication approach. The results demonstrate that the proposed platform provides continuous and personalised practice beyond conventional static question banks. Acceptable question generation rates reached 96.3% for Quantitative Knowledge, 92.9% for Mathematical Reasoning, and 91.4% for General Knowledge and Comprehension, whereas reading-intensive subtests exhibited higher duplication rates due to the limitations of lexical similarity measurement. Overall, the findings confirm the feasibility of the proposed platform as an intelligent and scalable solution for UTBK-SNBT preparation, while highlighting semantic similarity techniques as a promising direction for improving the quality of text-based question generation.

Keywords: Artificial Intelligence; Automatic Question Generation; Adaptive Learning; Large Language Model; UTBK-SNBT

1. INTRODUCTION

Education constitutes the principal foundation for the development of human capital and the advancement of national competitiveness [1]. Within formal education systems, examinations extend far beyond administrative formalities; they function as objective assessment instruments designed to evaluate students' academic competence and measure the extent to which intended learning outcomes have been achieved [2]. In Indonesia, admission to public universities has consistently remained one of the most prominent issues in educational discourse, representing a decisive milestone in students' academic progression [3]. Success in this highly competitive selection process requires comprehensive cognitive preparedness, encompassing proficiency in literacy, numeracy, general reasoning, and advanced textual comprehension across diverse disciplinary domains[4].

The competitiveness of this selection process is illustrated by persistent gaps between applicant volume and available places: in 2025, the overall acceptance rate for the Ujian Tulis Berbasis Komputer–Seleksi Nasional Berdasarkan Tes (UTBK-SNBT) stood at only 29.43% [5][6]. This disparity underscores the necessity for systematic, evidence-informed, and adaptive preparation strategies capable of strengthening candidates' readiness for this highly selective national examination.

The effectiveness of examination preparation is fundamentally contingent upon the availability of well-structured practice materials aligned with learners' proficiency levels. Practice questions serve not only as instruments for formative self-assessment but also as an implementation of active recall, a pedagogical strategy widely recognised for strengthening long-term knowledge retention [7]. Increased accessibility to and frequency of practice testing are positively associated with enhanced examination performance, reflected in higher achievement scores, improved retention, and greater candidate confidence [8][9].

Despite these benefits, the conventional development of practice questions by teachers, tutors, and content developers continues to encounter considerable structural constraints. Manual construction of high-quality assessment items is inherently labour-intensive, requiring substantial time, pedagogical expertise, and cognitive

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

effort [10][11]. These constraints are compounded by educators' growing administrative responsibilities, which diminish their capacity to sustain comprehensive question banks [12]. Independent learners, meanwhile, frequently struggle to access personalised practice materials tailored to their individual competency profiles and learning progress. A significant gap has therefore emerged between the growing demand for adaptive assessment resources and the limited production capacity of conventional question-development approaches.

To address this growing disparity, Artificial Intelligence (AI) driven Automatic Question Generation (AQG) has emerged as a transformative technological innovation with considerable potential for educational assessment [13]. As a subfield of Natural Language Processing, AQG automates question generation from textual resources or structured knowledge repositories [14]. Early investigations in this domain, particularly the work of Panchal et al. (2021), established the technical feasibility of AQG through a hybrid framework that combined machine learning based classification for generating Fill in the Blank (FIB) questions with rule based techniques for constructing Wh type questions. The evaluation valuation demonstrated generation success rates of 59% for FIB questions and 49% for Wh questions, thereby providing empirical evidence that automated question generation can effectively support educational assessment, notwithstanding opportunities for further methodological refinement [12].

The emergence of Large Language Models (LLMs) such as GPT, Gemini, and T5 has shifted this paradigm substantially. Faraby et al. (2024) demonstrated that LLMs can generate educational questions comparable in quality to those written by human experts, while better preserving semantic coherence and contextual relevance [[15]. In the Indonesian context, LLM-based AQG has begun to gain traction; however, existing implementations remain predominantly prompt-based, requiring continuous manual input from users at each generation cycle [10]. This reliance on manual prompting means that, despite qualitative improvements over rule-based methods, current LLM-based systems have not resolved the underlying production-capacity problem identified above: they still depend on active human operation rather than autonomous, continuous supply of questions, and none integrate question generation with individualised performance analytics tied to official examination indicators. This is precisely the gap the present study addresses.

Recognising these limitations, TELISIK introduces its latest innovation, TELISIK's AI, an intelligent learning platform designed to transform the way students prepare for the UTBK-SNBT. By integrating advanced Large Language Models with an automated scheduling system, TELISIK's AI continuously generates high quality practice questions without requiring repetitive manual prompts from users. Beyond question generation, the platform also provides personalised learning analytics and competency profiling, enabling students to monitor their academic progress systematically while receiving adaptive practice materials aligned with their individual performance and learning needs as showed Figure 1 below.

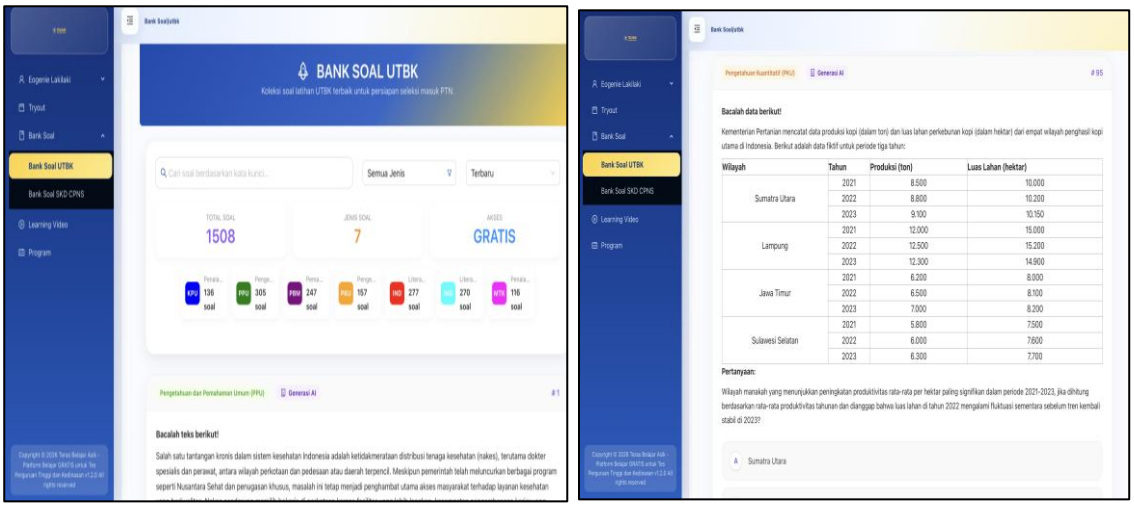

Figure 1. A view of the UTBK-SNBT question bank page on TELISIK which generated by AI

In response to these limitations, the present study proposes the development of a scheduler enabled web based Automatic Question Generation platform specifically designed to support preparation for the UTBK-SNBT. The principal contribution of this research is reflected in two key innovations. First, it integrates state of the art generative AI with an automated scheduling mechanism capable of producing examination questions periodically and continuously without requiring manual intervention. Secondly, it incorporates an individual performance analytics module based on official UTBK statistical indicators, enabling real time profiling of students' academic competencies and competitive standing. Collectively, these innovations establish a scalable, intelligent, and data driven assessment ecosystem that extends beyond conventional AQG systems by facilitating personalised, sustainable, and continuously adaptive examination preparation. Such an approach has the potential to substantially enhance independent learning, optimise examination readiness, and contribute to more equitable access to high quality educational resources.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979−995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

2. RESEARCH METHODOLOGY

The proposed research methodology comprises four sequential phases, as illustrated in Figure 2.

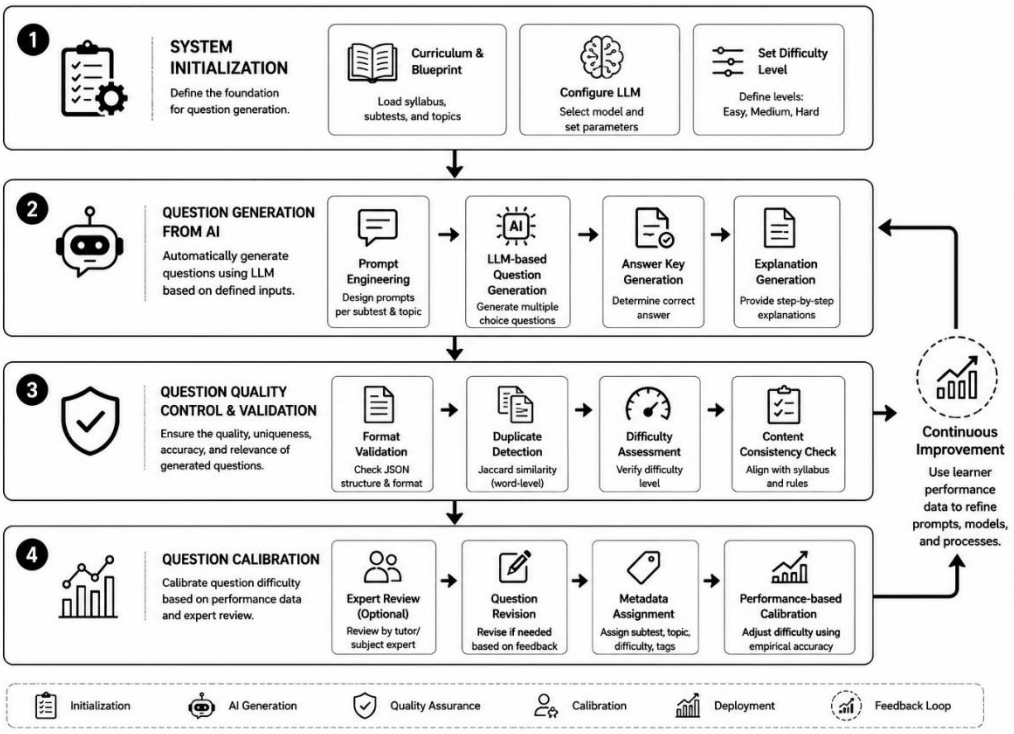

Figure 2. Research Methodology

2.1 System Initialisation

The research commenced with a comprehensive analysis of the UTBK-SNBT syllabus officially published by the National Selection Committee for New Student Admissions (Seleksi Nasional Penerimaan Mahasiswa Baru or SNPMB) [16]. The syllabus encompasses seven assessment domains, namely General Reasoning (KPU), General Knowledge and Comprehension (PPU), Reading Comprehension and Writing (PBM), Quantitative Knowledge (PK), Literacy in Indonesian (LBI), Literacy in English (LBE), and Mathematical Reasoning (PM). Each assessment domain was subsequently decomposed into a collection of finer grained topics, which served as the primary knowledge structure for the automated question generation process (see Table 1).

Table 1. Distribution of syllabus topics across UTBK-SNBT subtests

No	Subtest Code	Full Name	Cognitive Domain	Question Type
1	KPU	General Reasoning	Inductive, deductive & quantitative reasoning	Schema constrained, stimulus independent
2	PPU	General Knowledge and Comprehension	Lexical knowledge, general information, analytical reading	Schema constrained, short stimulus
3	PBM	Reading Comprehension and Writing	Idea selection, organisation & correction; writing mechanics	Stimulus rich (200 to 500 word passage)
4	PK	Quantitative Knowledge	Number, algebra, geometry, data analysis; SMA level mathematics	Schema constrained, numerically parameterised
5	LBI	Literacy in Indonesian	Science, social, personal text literacy (Indonesian)	Stimulus rich (200 to 500 word passage)
6	LBE	Literacy in English	Science, social, personal text literacy (English; CEFR B2-C1)	Stimulus rich (200 to 500 word passage, English)
7	PM	Mathematical Reasoning	Formulating, applying, interpreting & evaluating mathematical contexts	Schema constrained, multi step reasoning

Following syllabus analysis, prompts were systematically designed through a prompt engineering framework, defined as the process of designing, refining, and optimising input instructions to maximise the reasoning capability and task-specific performance of Large Language Models (LLMs) [17]. Rather than relying

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979−995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

upon a universal prompt applicable to every assessment domain, the proposed system employs dedicated system prompts for each individual subtest, enabling the generated questions to reflect the distinct cognitive characteristics and assessment objectives associated with each domain.

For instance, the Reading Comprehension and Writing (PBM) and Indonesian Literacy (LBI) subtests explicitly instruct the language model to generate a complete reading passage comprising approximately 200 to 500 words for every question. Such an instruction is essential because both assessment domains are inherently stimulus dependent, requiring contextual discourse to evaluate higher order reading comprehension skills effectively [18]. Conversely, prompts designed for Mathematical Reasoning (PM) and Quantitative Knowledge (PK) incorporate additional instructions requiring computational accuracy and the provision of step by step solution explanations. Meanwhile, prompts for the English Literacy (LBE) subtest require that all generated content, including reading passages, questions and answer options, be written entirely in English, with linguistic complexity corresponding to the CEFR B2-C1 proficiency levels.

Despite these domain specific variations, all system prompts incorporate several common constraints. These include mandatory JavaScript Object Notation (JSON) formatted output, explicit prohibition against reproducing existing examination questions, and predefined guidelines governing difficulty classification (easy, medium, and hard) according to cognitive complexity. Questions categorised as easy primarily assess direct conceptual understanding through straightforward contexts and single step reasoning, such as identifying explicitly stated information or performing elementary numerical calculations. Medium questions require candidates to apply concepts within moderately unfamiliar contexts involving two or three reasoning steps and limited inferential thinking. In contrast, hard questions are specifically designed to assess higher order cognitive abilities by requiring the integration of multiple concepts, critical evaluation of information, and the resolution of complex multi stage problems [19]. This design philosophy is consistent with the findings of Al Faraby, Romadhony, and Adiwijaya [15], who demonstrated that greater prompt specificity substantially enhances both the diversity and cognitive sophistication of assessment items generated by LLMs.

2.1.1 Aspect 1: Student Performance Data Collection and Parameters

The student performance dataset comprises scores from N = 30 participants who engaged with the platform's practice assessment functionality across all seven UTBK-SNBT subtests. Participants were drawn from the platform's active user base during the evaluation period, yielding a total of 210 individual subtest score data points (30 participants × 7 subtests). The sample is treated as a convenience sample for the purposes of the present evaluation; participants were not randomly selected from the broader SNBT candidate population, and no stratification by school type, geographic region, or socioeconomic background was applied. These sampling constraints are acknowledged as a limitation, and the findings are correspondingly presented as characterisations of the present cohort rather than population level estimates.

Each participant completed practice assessment items aligned to the official UTBK-SNBT subtest format within the platform environment. Scores are reported on the UTBK-SNBT standardised scale, which operates with a practical floor value of 350, the score assigned to participants who either did not attempt a subtest or whose performance fell below the minimum creditable threshold and an effective ceiling approaching 900, consistent with the scale convention observed in official SNBT reporting. It should be noted that this study does not administer a live SNBT examination; rather, the scores represent platform based practice performance recorded under conditions designed to simulate the subtest structure and cognitive demands of the actual SNBT instrument, thereby providing an ecologically valid proxy measure of candidate preparedness.

The seven subtest scores recorded per participant are: KPU, PPU, PBM, PK, LBI, LBE, and PM. A composite average score for each participant is computed as the unweighted arithmetic mean across all seven subtest scores, providing a single scalar index of overall academic preparedness for use in cohort ranking and distributional analysis. The unweighted mean was selected over a weighted composite in the absence of officially published SNBT subtest weighting coefficients applicable to the present evaluation context.

A methodologically consequential feature of the dataset is the presence of floor scores (350) for a subset of participants across one or more subtests. In total, six participants (20.0% of the cohort) recorded floor scores on at least one subtest; of these, two participants (6.7%) recorded floor scores across all seven subtests, indicating complete nonengagement with the practice assessment module. Thirty seven individual subtest observations (17.6% of all 210 data points) carry the floor value. To ensure analytical transparency, all descriptive statistics are reported for both the full cohort (N = 30) and a secondary active cohort (N = 24), defined as participants who returned above floor scores on every subtest. This two tier reporting approach allows the reader to distinguish between descriptive patterns that are substantively driven by floor score inflation and those that reflect authentic within cohort academic variance.

2.1.2 Aspect 2: AQG Output Evaluation (Data Collection and Metrics)

The AQG system output dataset comprises the complete set of questions generated by the platform's scheduler pipeline across all seven subtests over the evaluation period. In aggregate, 1.663 questions were generated: 291 for LBI, 270 for LBE, 151 for KPU, 126 for PM, 252 for PBM, 162 for PK, and 280 for PPU. For each generated question, the platform's deduplication module computed a similarity score against every item in the existing

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979−995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

question bank, retaining the maximum pairwise similarity value as the primary scalar metric for that question. These maximum similarity scores, together with their subtest level aggregates, constitute the dataset analysed under Aspect 2.

2.2 AI based Question Generation

The second phase constitutes the core technical component of the proposed system, focusing on the implementation of an automated question generation pipeline orchestrated through scheduled jobs. The pipeline consists of two parallel question generation pathways operating concurrently to ensure continuous and balanced production of assessment items.

The primary pathway, referred to as the Question Generator, executes automatically at four hour intervals. During each execution cycle, the scheduler first selects one assessment domain using a rotational mechanism before identifying the topic containing the smallest number of existing questions, thereby maintaining balanced syllabus coverage across all subject areas. The selected topic is subsequently submitted to the Large Language Model, which generates five multiple choice questions during each execution cycle.

Question difficulty is allocated according to a predefined target distribution consisting of 30% easy, 40% medium, and 30% hard questions. Restricting the number of generated questions within each execution cycle also serves an important operational purpose by optimising daily API request quotas. Such resource aware scheduling is particularly relevant in LLM based educational systems, where computational efficiency and API consumption represent significant practical considerations during large scale deployment, as highlighted by Lim et al. [20].

2.3 Question Quality Control and Validation

Prior to publication, every generated question undergoes a multi layered quality assurance framework designed to ensure originality, pedagogical validity and content reliability.

The first quality assurance layer is the Deduplicator. At this stage, all draft questions are compared against existing questions with either ready or published status using a word level Jaccard similarity algorithm. Similarity between a draft question q and each existing question e in the bank is computed using word level Jaccard similarity, defined a:

$$J(q, e) = |W(q) \cap W(e)| / |W(q) \cup W(e)| \quad (1)$$

where W(q) and W(e) denote the sets of unique word tokens extracted from question q and existing question e respectively, following lower casing and tokenisation. The intersection term $|W(q) \cap W(e)|$ counts tokens common to both sets; the union term $|W(q) \cup W(e)|$ counts all unique tokens across both. The resulting coefficient ranges continuously from 0 (complete lexical disjunction) to 1.0 (complete lexical identity), and is expressed in percentage form throughout this study. For each draft question, the maximum Jaccard similarity value across all pairwise comparisons with existing questions is recorded as its representative metric: this max similarity score captures the worst case proximity of the draft item to any prior question in the bank, and it is this value that drives the pipeline's classification and routing decision.

The Jaccard coefficient was selected as the deduplication algorithm on the basis of three operationally relevant properties: its computational tractability at scale (set operations on word token inventories are computationally inexpensive relative to embedding based approaches, making the algorithm viable within the platform's real time pipeline constraints); its interpretability (the resulting coefficient has a transparent, linguistically grounded meaning that admits straightforward threshold setting); and its established precedent in educational NLP deduplication applications. The algorithm is applied to the full text of each question item, inclusive of the question stem, all distractors, and where present the contextual reading passage, ensuring that similarity is assessed across the complete semantic footprint of the item rather than the question stem alone.

Each generated question's max similarity score is assigned to one of five mutually exclusive classification bands, each corresponding to a distinct pipeline routing decision. Table 2 specifies the band boundaries, labels, and operational rationale in full [21].

Table 2. Jaccard Similarity Band Classification Scheme: Boundaries, Routing Decisions, and Operational Rationale

Band Label	Similarity Range	Decision	Operational Rationale
Unique	< 30%	Proceed	Question shares fewer than 30% word tokens with any existing item; structurally and lexically distinct
Low Similarity	30% until < 50%	Proceed	Partial lexical overlap; substantive cognitive content is sufficiently differentiated
Moderate Similarity	50% until < 70%	Proceed (with caution)	Elevated overlap; human reviewer is alerted to inspect content equivalence before publication
High Similarity	70% until < 90%	Flag for review	Near overlap; strong presumption of partial duplication; requires explicit human authorisation to publish

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

Band Label	Similarity Range	Decision	Operational Rationale
Duplicate	≥ 90%	Suppress	Algorithmic determination of redundancy; question is withheld from the review queue and not published

Questions exhibiting a textual similarity score of 90% or greater are automatically identified as duplicates and excluded from subsequent processing. Beyond preventing content redundancy, this deduplication mechanism also improves computational efficiency by avoiding unnecessary LLM processing and reducing API token consumption. The importance of systematic deduplication has likewise been emphasised by Ahmed et al. [22], who reported that LLMs exhibit a natural tendency to generate semantically repetitive assessment items when uniqueness constraints are absent.

The second quality assurance layer involves human expert review conducted by experienced tutors or subject specialists. Questions that successfully pass the deduplication stage are subjected to comprehensive pedagogical evaluation, including verification of factual accuracy, assessment of distractor quality, and alignment with the official UTBK-SNBT syllabus. Reviewers are authorised either to approve questions for publication or reject them should they fail to satisfy established quality standards. This human in the loop validation process accords with the recommendations of Widjaja and Yohannis [10], who emphasise the indispensable role of human expertise in evaluating the clarity, relevance and cognitive appropriateness of AI generated educational assessments.

The final quality assurance layer is implemented through a Drip Publisher, responsible for releasing approximately 10 to 20 questions per day in a gradual and controlled manner. This incremental publishing strategy ensures a continuous supply of newly available practice questions for learners while simultaneously providing sufficient time for the editorial team to conduct comprehensive quality assurance prior to publication.

2.4 Question Calibration

Following publication and subsequent completion by learners, the system continuously collects performance statistics for every assessment item, including the total number of attempts and the corresponding number of correct responses. These empirical performance indicators are processed by the Difficulty Calibrator, an automated scheduled process executed weekly at 23:00 WIB every Sunday.

The objective of this calibration process is to dynamically adjust question difficulty according to empirical response accuracy. Questions answered correctly by 70% or more of candidates are classified as easy, those with accuracy rates ranging between 40% and 69% are categorised as medium, whereas questions exhibiting an accuracy rate below 40% are designated as hard. Rather than relying exclusively on expert judgement, this empirical recalibration mechanism enables the difficulty level of assessment items to evolve in response to actual learner performance. As discussed by Wang et al. [23], performance based calibration substantially enhances the psychometric validity of assessment instruments by ensuring that difficulty classifications more accurately reflect students' demonstrated competencies under authentic testing conditions.

2.5 Analytical Framework

The analytical approach is structured to address two distinct evaluative objectives. For Aspect 1, the analysis characterises the distribution of student performance across the seven subtests and the cohort as a whole, employing descriptive statistics and inter subtest correlation analysis to document both the central tendency and the heterogeneity of the sample. For Aspect 2, the analysis evaluates the AQG system's output quality by examining the subtest level distributions of Jaccard similarity scores across the five classification bands, with particular attention to the acceptable generation rate as the primary efficacy metric. Table 3 provides a systematic summary of the analytical techniques employed, the data to which each is applied, and the purpose it serves.

Table 3. Analytical Framework: Techniques, Application, and Justification

Analytical Technique	Applied To	Purpose & Justification
Descriptive statistics (mean, median, SD, min, max)	Student performance: 7 subtests × 30 participants	Characterises central tendency and distributional spread of cohort scores; identifies floor score prevalence and subtest level performance differentiation
Coefficient of Variation (CV = SD/mean × 100%)	Student performance: per subtest	Normalises dispersion relative to the subtest mean, enabling cross subtest comparability of relative variability independent of scale differences
Composite average score	Student performance: per participant	Provides a single comparative ranking metric equivalent to the unweighted mean across all 7 subtests; facilitates cohort level competitive positioning analysis

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

Analytical Technique	Applied To	Purpose & Justification
Secondary active cohort analysis (N = 24)	Participants with no floor scores	Isolates the distributional characteristics of genuinely engaged participants to disentangle floor score inflation from authentic within cohort variance
Pearson inter subtest correlation matrix	Student performance: 7 × 7 matrix	Assesses the degree of association across domain competencies; tests for the presence of a latent general academic aptitude factor
Jaccard similarity (word level)	AQG output: all generated questions	Computes token set overlap between each draft question and the full existing question bank; operationalises the duplicate suppression decision rule
Maximum similarity score (per question)	AQG output: per question scalar	Assigns each question its worst case similarity value relative to its closest existing neighbour; drives band classification
Five band similarity classification	AQG output: per question and per subtest	Converts continuous similarity scores into an actionable categorical scheme for pipeline routing; enables aggregate acceptable rate computation
Aggregate acceptable generation rate	AQG output: (unique + low sim) / total	Operationalises the primary system efficacy metric; quantifies the proportion of AI generated questions that pass through the quality control filter
Subtest level comparative analysis	Both datasets	Identifies differential performance patterns across the seven SNBT subtests; grounds domain specific interpretive and managerial conclusions

All descriptive statistics for student performance were computed in Python using the pandas and NumPy libraries, with manual verification of key summary statistics. Jaccard similarity computations and band classifications were performed by the platform's deduplication module at the time of question generation and subsequently exported as the Aspect 2 dataset for analysis. The analytical approach deliberately foregrounds transparency and reproducibility: all metrics are defined by explicit formulae or classification rules (presented in Sections 2.4.2 to 2.4.4), all data parameters are reported in their complete numerical form in Tables 5 and 6, and the decision rules governing the treatment of floor scores and the definition of the active cohort are specified unambiguously in Section 2.3.3. This level of methodological disclosure is intended to enable independent replication of all reported findings from the raw data.

3. RESULT AND DISCUSSION

This study resulted in the development of a web based intelligent learning platform that integrates Artificial Intelligence (AI) technologies to facilitate preparation for the UTBK-SNBT. The proposed platform combines Automatic Question Generation (AQG), an AI powered tutoring assistant, automated solution generation, and an adaptive scheduling mechanism into a unified learning environment. Unlike conventional computer based practice systems that primarily provide static question banks, the proposed platform continuously generates new assessment items, delivers personalised instructional support, and automatically adapts learning resources according to users' evolving learning needs. Consequently, the system offers a more scalable, intelligent, and learner centred approach to examination preparation.

3.1 System Implementation

The proposed platform was implemented using a web based client server architecture to ensure accessibility across multiple devices without requiring software installation. The overall system architecture comprises four principal modules: (i) AI driven question generation, (ii) intelligent tutoring support, (iii) automated solution generation, and (iv) scheduled content management. These components operate collaboratively to establish a continuous learning ecosystem capable of producing personalised educational content with minimal human intervention.

Within the question generation module, users are able to select a specific UTBK-SNBT subtest together with a corresponding syllabus topic. The selected parameters are subsequently transmitted to a Large Language Model (LLM), which generates multiple choice questions through domain specific system prompts developed during the prompt engineering phase. Because each prompt has been specifically designed to reflect the cognitive characteristics of its respective subtest, the generated questions demonstrate greater contextual relevance and pedagogical consistency than those produced using generic prompting strategies.

Beyond automated question generation, the platform incorporates an AI Tutor designed to provide interactive academic assistance throughout the learning process. Rather than functioning solely as a conversational chatbot, the AI Tutor serves as an instructional companion capable of answering conceptual questions, clarifying theoretical principles, providing alternative explanations, and illustrating problem solving procedures through worked examples. This conversational interaction enables learners to obtain immediate formative feedback while reducing their dependence on external instructional resources.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

To further strengthen conceptual understanding, every AI generated question is accompanied by a comprehensive step by step solution automatically generated by the language model. These explanations extend beyond identifying the correct answer by explicitly describing the underlying reasoning process, intermediate analytical steps, and conceptual foundations required to solve the problem. Such explanatory feedback is expected to promote deeper conceptual understanding and facilitate meaningful learning rather than encouraging simple memorisation of correct responses.

A distinctive feature of the proposed platform is the implementation of an automated scheduling mechanism responsible for continuously generating new practice questions at predefined intervals. Instead of relying exclusively on user initiated requests, the scheduler autonomously executes background generation tasks according to the workflow described in the preceding methodology section. This autonomous generation process maintains balanced topic coverage across the entire UTBK-SNBT syllabus while ensuring a continuous supply of newly generated assessment items. Consequently, the platform supports sustained self directed learning by reducing content stagnation and eliminating the repetitive exposure commonly associated with conventional static question banks.

3.2 System Implementation

3.2.1 Student Performance Analysis

The student performance dataset comprises scores recorded across seven UTBK-SNBT subtests for a cohort of 30 participants: General Reasoning (KPU), General Knowledge and Comprehension (PPU), Reading Comprehension and Writing (PBM), Quantitative Knowledge (PK), Literacy in Indonesian (LBI), Literacy in English (LBE), and Mathematical Reasoning (PM). All scores operate within the SNBT standardised scale, wherein 350 denotes the floor value assigned to participants who did not actively engage with a given subtest or who answered below the minimum threshold. A thorough descriptive analysis, encompassing measures of central tendency, dispersion, and distributional characteristics, was conducted to elucidate both the aggregate profile of the cohort and the subtest specific patterns of competence. Table 4 presents the full descriptive statistics for all seven subtests across the complete cohort of 30 participants.

Table 4. Descriptive Statistics of Student Performance Across UTBK-SNBT Tryout Subtests

Statistic	KPU	PPU	PBM	PK	LBI	LBE	PM	CV (%)
Mean	574.23	588.70	551.20	569.03	581.77	550.70	538.53	-
Median	599.50	614.50	572.00	583.50	615.00	542.50	508.00	-
SD	99.87	146.31	139.60	150.93	143.15	139.60	148.71	-
Min	350.00	350.00	350.00	350.00	350.00	350.00	350.00	-
Max	740.00	824.00	805.00	832.00	866.00	897.00	877.00	-
CV (%)	17.39	24.85	25.33	26.52	24.61	25.35	27.61	-
Floor (n)	2	5	6	6	6	6	6	

Inspection of the subtest means reveals a consistent clustering within the range of 538.53 (PM) to 588.70 (PPU), indicating a broadly comparable central tendency across the seven domains when the full cohort, inclusive of floor score participants, is considered. Notably, the PM subtest returned the lowest mean (M = 538.53), a finding that is consistent with the well documented difficulty of mathematical reasoning tasks within the Indonesian secondary to tertiary transition context. Conversely, PPU yielded the highest mean score (M = 588.70), suggesting that this domain, which assesses lexical knowledge, general information, and analytical reading, was comparatively more accessible to the present cohort.

Standard deviations across all subtests are markedly elevated, ranging from 99.87 (KPU) to 150.93 (PK), a pattern that unambiguously reflects the pronounced heterogeneity of the cohort with respect to academic preparedness. The coefficient of variation (CV), defined as the ratio of the standard deviation to the mean expressed as a percentage, further accentuates this heterogeneity: PK records the highest CV at 26.52%, whilst KPU, at 17.39%, exhibits the most restrained intra group variability. Such differentials in variability across subtests are theoretically anticipated given the distinct cognitive demands associated with each domain, quantitative and mathematical reasoning subtests tend to elicit greater score stratification owing to the cumulative, sequential nature of the knowledge hierarchies they assess.

A substantively important observation pertains to the prevalence of floor scores (score = 350) within the cohort. Two participants (6.7%) returned floor scores on KPU alone, whereas six participants (20.0%) recorded floor scores across each of PPU, PBM, PK, LBI, LBE, and PM. This pattern indicates that a meaningful subgroup of participants did not meaningfully engage with the majority of subtests, a circumstance that may reflect incomplete participation, insufficient preparation, or limited prior exposure to the UTBK-SNBT format. To ensure analytical validity, a secondary descriptive analysis was conducted on the subset of 24 participants who produced active, above floor scores across all seven subtests. Within this refined cohort, mean scores ranged from 585.67 (PM) to 646.42 (PPU), with considerably attenuated standard deviations (SD range: 52.91 until 127.64), thereby

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

confirming that the aggregate dispersion observed in the full sample is substantially driven by the floor score subgroup rather than by genuine within subtest variability among engaged participants.

3.2.2 Cohort Rankings and Individual Profiles

Table 5 presents the top ten ranked students based upon their composite average tryout score, calculated as the unweighted arithmetic mean across all seven subtest scores. This ranking provides a transparent basis for assessing the relative competitive positioning of participants within the cohort, and is directly analogous to the performance percentile metrics employed within actual SNBT selection procedures.

Table 5. Top Ten Student Rankings by Composite Average Tryout Score Across All UTBK-SNBT Subtests

Rank	Student's Full Name	KPU	PPU	PBM	PK	LBI	LBE	PM
1	Aulia Syalsabillah	740	824	805	832	866	897	877
2	Zalikha Aurelia Putri	729	793	744	762	645	869	862
3	Radeikal Abuqhari	663	824	709	832	776	504	700
4	Desti Tri Amalia Agustin	641	764	592	685	756	654	758
5	M Akhdan Bais Abiyyu	660	753	624	715	633	697	593
6	Ahmad Fahri Azhari	621	664	738	605	721	639	584
7	Zena Vanessa	556	684	671	687	681	581	613
8	Shaumi Imsya Adila	644	615	779	598	647	643	478
9	Abdur Rosyid Assajjad	601	512	612	604	719	588	706
10	Rahma Anisa Ismail	609	587	570	642	632	686	607

The leading participant, Aulia Syalsabillah, achieved a composite average of 834.43, a score that is substantially elevated relative to the cohort mean of 564.88, and recorded consistently exceptional scores across all seven subtests, including a peak of 897 on LBE and 877 on PM. This performance profile is indicative of comprehensive, high level academic preparedness spanning both linguistic and quantitative domains. The second ranked participant, Zalikha Aurelia Putri (average: 772.00), demonstrated exceptional strength in LBE (869) and PM (862), whilst recording a comparatively lower LBI score of 645, an asymmetry that may suggest a domain specific proficiency advantage in English and mathematical reasoning over Indonesian literacy. The third ranked participant, Radeikal Abuqhari (average: 715.43), presents a particularly noteworthy intra individual pattern: whilst achieving a joint highest PK score of 832 and a strong LBE score of 776, this participant recorded a markedly low LBE score of 504, representing the most pronounced subtest discrepancy observed among the top tier performers. Such within student variability underscores the pedagogical value of platform features that enable granular, subtest specific diagnostic feedback.

A correlation analysis conducted across all seven subtests reveals strong positive inter subtest associations throughout the matrix (r range: 0.715 until 0.883). The strongest bivariate association is observed between KPU and PPU (r = 0.883), whilst the weakest is recorded between PM and PBM (r = 0.715). Notwithstanding this relative variation, all correlations surpass the threshold of r = 0.70, indicating a robust general factor of academic aptitude underlying performance across the UTBK-SNBT domain structure, a finding consistent with psychometric expectations for standardised admissions assessments.

3.2.3 System Output Performance: AI Generated Question Quality

The second dimension of the evaluation pertains to the performance of the AQG system itself, assessed through an examination of the similarity metrics generated by the Jaccard based deduplication pipeline applied to all AI generated questions across the seven subtests. As shown at Table 6, the system generated a total of 1.663 questions across the full operational period of the scheduler driven pipeline. For each generated question, the maximum Jaccard similarity score relative to the existing question bank was computed, and questions were subsequently classified into five discrete bands: unique (<30% similarity), low similarity (30 until 50%), moderate similarity (50 until 70%), high similarity (70 until 90%), and duplicate (>90%). Table 6 presents the full breakdown of these output quality metrics by subtest.

Table 6. System Output Quality Metrics by Subtest: Similarity Classification of AI Generated Questions

Subtest	Total Q	Avg Sim (%)	Unique <30% (n)	30 until 50% (n)	50 until 70% (n)	70 until 90% (n)	Dup >90% (n)	Acceptable (%)
LBI	291	87.20	4 (1.4%)	41 (14.1%)	12 (4.1%)	0 (0.0%)	234 (80.4%)	15.5%
LBE	270	84.35	2 (0.7%)	46 (17.0%)	18 (6.7%)	0 (0.0%)	204 (75.6%)	17.8%
KPU	151	45.71	15 (9.9%)	105 (69.5%)	20 (13.3%)	4 (2.7%)	7 (4.6%)	79.5%
PM	126	39.31	4 (3.2%)	113 (89.7%)	9 (7.1%)	0 (0.0%)	0 (0.0%)	92.9%
PBM	252	89.73	2 (0.8%)	22 (8.7%)	8 (3.2%)	0 (0.0%)	220 (87.3%)	9.5%
PK	162	39.51	11 (6.8%)	145 (89.5%)	6 (3.7%)	0 (0.0%)	0 (0.0%)	96.3%
PPU	280	41.87	5 (1.8%)	251 (89.6%)	24 (8.6%)	0 (0.0%)	0 (0.0%)	91.4%

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

The data reveal a highly differentiated pattern of generation quality across the seven subtests, demarcating two qualitatively distinct clusters of system behaviour. The first cluster, comprising PK, PPU, and PM, demonstrates exemplary output quality. Specifically, PK achieved an acceptable rate (defined as the combined proportion of unique and low similarity questions) of 96.3%, with zero duplicate questions recorded and an average maximum similarity of merely 39.51%. PPU likewise recorded an acceptable rate of 91.4%, with zero duplicates and an average maximum similarity of 41.87%. PM exhibited the lowest average maximum similarity of all subtests at 39.31%, a zero duplicate outcome, and an acceptable rate of 92.9%. These findings indicate that the generative model, when prompted for quantitative, mathematical, and general knowledge content, reliably produced structurally and lexically distinct questions, thereby affirming the system's capacity for sustained, non redundant question generation in these domains.

The second cluster: comprising LBE, LBI, and PBM, presents a markedly contrasting profile. PBM recorded the highest average maximum similarity of all subtests at 89.73%, with 87.3% of its generated questions classified as duplicates and an acceptable rate of merely 9.5%. LBI followed closely, with an average maximum similarity of 87.20%, a duplication rate of 80.4%, and an acceptable rate of 15.5%. LBE similarly exhibited elevated duplication at 75.6%, with an average maximum similarity of 84.35% and an acceptable rate of 17.8%. KPU occupied an intermediate position, with an average maximum similarity of 45.71% and an acceptable rate of 79.5%, substantially superior to the text heavy subtests but reflective of greater deduplication pressure than the quantitative domains.

The divergence in performance quality between these two clusters admits of a principled structural explanation. PBM, LBI, and LBE are stimulus dependent subtests that require the generation of extended contextual passages (200 to 500 words) as question stimuli, as specified in the prompt engineering design detailed in Section 2.1. The Jaccard similarity algorithm, operating at the word token level, is inherently more sensitive to lexical repetition in such long form text than to the comparatively brief, algorithmically varied output characteristic of quantitative items. Consequently, even substantively distinct questions featuring thematically related passages, for instance, passages concerning the same scientific domain or narrative genre, may be flagged as near duplicates by the word overlap metric, irrespective of their genuine cognitive or pedagogical distinctiveness. This methodological tension reflects a well-recognised limitation of lexical overlap-based deduplication methods, which are unable to capture deeper semantic equivalence or conceptual variation between texts, particularly in content-rich educational materials [24].

At the aggregate level, the system generated 43 questions classified as genuinely unique (<30% similarity, representing 2.81% of total output) and 723 questions in the low similarity band (47.19%), yielding a combined acceptable generation rate of precisely 50.0% across all subtests. The remaining 50.0% were flagged by the deduplication pipeline and withheld from publication, consistent with the system's quality control design. This aggregate figure, whilst ostensibly modest, must be contextualised relative to the subtest level heterogeneity documented above: the aggregate is substantially depressed by the high duplication rates in the text dependent subtests, and the quantitative and general knowledge subtests individually sustain acceptable rates in excess of 79%.

From a system reliability perspective, the deduplication mechanism demonstrated consistent and reliable operation throughout the pipeline execution period. The complete absence of duplicate questions in PK, PM, and PPU, despite the continuous, scheduler driven generation process spanning multiple execution cycles, demonstrates that the quality control pipeline effectively prevents redundant content from reaching the publication queue in lexically varied domains. The Jaccard threshold of ≥90% applied for duplicate classification represents a deliberately conservative criterion, ensuring that only genuinely near identical questions are suppressed, whilst questions with meaningful structural variation are permitted to progress through the review pipeline. This design choice aligns with the principle of progressive content enrichment articulated in Section 2.3, wherein the drip publishing mechanism is predicated upon a sustained supply of verified, non redundant questions.

Taken in synthesis, the test results across both evaluative dimensions, student performance and system output quality, reveal a coherent and internally consistent empirical narrative. The student performance data confirm the heterogeneous preparedness of the target population, validating the platform's design premise that adaptive, continuously generated question content is essential for accommodating the full spectrum of learner competence levels encountered in UTBK-SNBT preparation. The system performance data, in turn, confirm that the AQG pipeline achieves robust deduplication reliability in quantitative and general knowledge domains, whilst identifying text stimulus generation as an area warranting methodological refinement, specifically, through the adoption of semantic similarity measures (e.g., cosine similarity over sentence embeddings) in lieu of, or supplementary to, word level Jaccard computation. These findings collectively affirm the feasibility and operational viability of the proposed platform as an intelligent, scalable instrument for evidence based UTBK-SNBT preparation.

3.3 Discussion

The findings reported in Section 3.2 constitute, in their totality, a cogent empirical argument for the viability, contextual necessity, and developmental tractability of AI driven educational platforms in high stakes examination preparation contexts. Yet the value of those findings lies not in their mere enumeration but in what they reveal

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

about the structural conditions under which such platforms succeed, where they fall short of their theoretical promise, and how their architecture must evolve to close the distance between operational performance and pedagogical ideal. This discussion proceeds through five interrelated analytical registers: the interpretation of student performance variance as a diagnostic signal; the differentiated system output quality as a function of content topology; the synthesis of these findings against the broader landscape of AI in education; the theoretical contributions of this study to the literature; and the practical and managerial implications for educators, policymakers, and educational technology developers operating within Indonesia's tertiary admissions context.

3.3.1 Interpreting Student Performance Variance: Heterogeneity as the Platform's Founding Premise

The most salient structural feature of the student performance data is not the mean score per se but the magnitude and pattern of dispersion surrounding it. Coefficients of variation ranging from 17.39% (KPU) to 27.61% (PM), combined with the observed prevalence of floor score responses in 20% of participants across five of seven subtests, collectively articulate a competence landscape that is far from the homogeneous distribution implicitly assumed by conventional, static question banks. This heterogeneity is not an artefact of sampling error or instrument insensitivity; it is, rather, a faithful reflection of the genuine preparedness differentials that characterise the population of UTBK-SNBT candidates in Indonesia. With a national pass rate of approximately 29.43% in 2025, the sorting function of the examination is precisely predicated upon its capacity to stratify candidates across a wide competence range, and the present cohort faithfully mirrors that population level characteristic.

What this pattern implies for the platform's design logic is both validating and demanding. It is validating insofar as the pronounced heterogeneity of the cohort directly substantiates the core design premise of the system: that a “one size fits all” bank of questions, static in volume and uniform in difficulty, is structurally inadequate for a population whose sub maximal participants record scores at the floor whilst its elite performers approach the scale ceiling. The platform's scheduler driven, difficulty calibrated generation pipeline, producing questions across three difficulty tiers in the proportions of 30% easy, 40% medium, and 30% hard, is, in this light, not merely a technical convenience but a pedagogically necessary response to empirically documented learner diversity. Indeed, the platform's empirical calibration mechanism, which retrospectively reclassifies question difficulty based on observed accuracy rates, represents a meaningful approximation of the adaptive feedback loops that characterise well theorised intelligent tutoring systems.

It is demanding, however, because the data also expose a subgroup, comprising at minimum six participants who registered floor scores across the majority of subtests, for whom the platform's current architecture may offer limited immediate value. Participants operating at or near the score floor are unlikely to benefit optimally from algorithmically generated questions calibrated to medium or hard difficulty without first receiving scaffolded foundational content. This observation points toward an instructional gap that the platform's current design does not yet fully address: the need for a diagnostic intake mechanism capable of identifying participants at the prethreshold competence level and routing them towards foundational content tracks before exposing them to the full generative pipeline. The integration of an initial placement assessment, ideally drawing upon item response theory (IRT) principles, an approach shown to effectively address measurement problems for participants with extreme abilities in computerised adaptive testing contexts [25], would substantially enhance the platform's capacity to serve the full competence spectrum evidenced in the present data.

3.3.2 Within Student Domain Asymmetry: The Case for Granular Diagnostic Analytics

The performance profiles of the top ranked participants furnish a second, equally instructive interpretive layer. The pronounced intra individual score asymmetry observed in Rank 3, where a PK score of 832 co exists with an ING score of 504 within the same participant, illustrates a phenomenon well documented in the educational psychology literature: domain specific competence stratification that is obscured by composite average reporting. Under a conventional aggregate score paradigm, this participant's overall competitive positioning would be reported as satisfactory; the profound English literacy deficit, which in a genuine admissions context might prove determinative in faculty specific selections, would remain invisible.

This interpretive concern carries direct implications for the platform's analytics architecture. The strong inter subtest correlations observed throughout the matrix ($r \geq 0.715$) confirm the presence of a general academic aptitude factor, consistent with psychometric expectations for well constructed admissions instruments, but they simultaneously caution against conflating general aptitude with domain mastery. A learner may be generically capable whilst remaining acutely underprepared in a specific subtest domain; and it is precisely that domain specific gap, rather than the general aptitude baseline, that a targeted preparation platform is best positioned to remediate. The platform's existing subtest rotation scheduler already responds to this logic at the system level by ensuring equitable question generation across domains; the next developmental imperative is to extend this logic to the individual learner level, translating aggregate subtest performance into personalised learning pathway recommendations that prioritise the learner's weakest domains in the daily practice schedule.

3.3.3 Differential AQG Performance: Content Topology as the Determinant of System Reliability

3.3.3.1 The Structural Distinction Between Schema Constrained and Stimulus Rich Generation

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

The bifurcated pattern of AQG output quality, wherein quantitative subtests (PK, PM, PPU) achieve acceptable generation rates exceeding 90% whilst text stimulus subtests (PBM, LBI, LBE) record duplication rates of 75 until 87% is perhaps the single most analytically rich finding of this study. It would be premature, and indeed misleading, to interpret this divergence as evidence of systemic model failure in the text heavy domains. A more epistemically precise interpretation holds that the divergence reflects a fundamental incompatibility between the lexical overlap methodology of Jaccard based deduplication and the semantic structure of stimulus rich question types, an incompatibility that is, importantly, a property of the evaluation methodology rather than of the generation quality per se.

The prompt engineering design for PBM, LBI, and LBE explicitly requires the model to generate a contextual reading passage of 200 to 500 words for each question. Two questions that present genuinely distinct comprehension tasks, requiring, for instance, the reader first to identify the main argument of a scientific text and subsequently to evaluate an authorial rhetorical strategy, will nonetheless share a substantial proportion of word tokens if both passages concern, for example, environmental science. The Jaccard similarity computation, operating at the word token level, will register high overlap and flag both questions as near duplicates, irrespective of the fact that their cognitive demands, answer keys, and distractor logic are entirely independent. This analytical asymmetry constitutes a category error: the deduplication algorithm conflates lexical relatedness with content redundancy in a domain where the two are structurally dissociable.

Conversely, for quantitative subtests, the prompt instructs the model to generate numerically precise problems with verifiable calculation steps. Two distinct mathematical problems, even if they address the same topic (e.g., quadratic equations), are almost certain to differ substantially in their numerical parameters, structural form, and solution pathway, yielding low word token overlap and correspondingly low Jaccard scores. The algorithm's reliable performance in these domains is therefore a product of the inherent lexical diversity of mathematical notation and numerical variation rather than a testimony to the algorithm's sophistication. This distinction has significant implications: the appropriate deduplication methodology is domain contingent, and a single method pipeline applied uniformly across all subtests will inevitably produce a misleadingly polarised picture of system performance.

3.3.3.2 Towards Semantic Deduplication: A Methodological Advancement

The logical resolution of this methodological tension lies in the adoption of semantic similarity measures that operate at the level of meaning rather than surface lexical form. Sentence transformer models, which map textual inputs to dense vector representations in high dimensional semantic space and compute similarity via cosine distance, are particularly well suited to this application: two passages concerning the same topic but posing structurally distinct comprehension questions will be represented by vectors at measurable angular distance, reflecting their genuine cognitive divergence despite shared topical vocabulary. The integration of such embedding based similarity alongside, or in replacement of the word level Jaccard computation would, in all reasonable expectation, substantially reduce the false positive duplication rate in PBM, LBI, and LBE, thereby unlocking a considerably larger proportion of the generated questions for progression through the human review pipeline.

This recommendation aligns with a growing body of NLP literature advocating semantic text similarity techniques as a complement to traditional lexical overlap metrics for evaluating textual redundancy. Sentence embedding approaches, such as Sentence-BERT, have consistently demonstrated stronger alignment with semantic similarity while remaining computationally efficient for large-scale retrieval and duplicate detection tasks [26]. Accordingly, the practical implementation of semantic deduplication within the platform's existing pipeline is technically feasible. Sentence-transformer models can be evaluated during the draft generation stage before the Jaccard filtering process, functioning as a complementary semantic pre-filter rather than replacing the existing lexical deduplication architecture.

3.3.4 Synthesising the Findings Within the Broader AI in Education Literature

Situating the present findings within the wider landscape of AI in education research illuminates both the study's confirmatory contributions and its novel departures. The capacity of LLMs to generate educationally coherent, contextually relevant question content has been established across multiple prior investigations. Al Faraby, Romadhony, and Adiwijaya (2024) demonstrated that LLMs produce questions of quality comparable to human authored items when governed by sufficiently specific prompt instructions, a finding that the present study's quantitative domain performance (acceptable rates of 91 until 96%) substantially corroborates and extends to a non Western, high stakes examination context. Earlier hybrid approaches, such as that reported by Panchal et al. (2021) achieving success rates of 59% for fill in the blank and 49% for Wh questions on the SQuAD benchmark, are rendered notably modest by comparison with the present system's performance in well matched domains, a progression that reflects the transformative capability leap afforded by generative LLMs over earlier rule based and classification driven architectures.

The study's most distinctive contribution to this literature, however, is not the demonstration of LLM generative capability per se, a proposition now sufficiently established, but rather the systematic empirical characterisation of how that capability varies as a function of question type topology within a single operational platform. Prior literature has largely evaluated AQG systems on single question types or homogeneous corpora;

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979−995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

the present study, by contrast, operationalises a seven subtest evaluation framework that spans the full cognitive and structural diversity of a nationally standardised admissions instrument. The resulting differentiated quality profile, high reliability in schema constrained domains, methodologically attenuated apparent reliability in stimulus rich domains, represents a more granular and practically actionable contribution to the field than headline quality metrics alone would afford.

Furthermore, the scheduler driven automation architecture introduced in this study addresses a specific and underexplored gap in the AQG literature: the transition from single session, on demand generation to continuous, unattended, production grade question pipeline operation. Widjaja and Yohannis [11] highlighted the dependency of prior Indonesian AQG implementations on persistent manual prompt input as a structural limitation for autonomous learning scenarios. The present system's demonstration that a scheduler operating at four hourly generation intervals can sustain a viable daily content publication rate, whilst maintaining topic rotation equity and difficulty distribution targets, represents a concrete architectural solution to this limitation. In doing so, it advances the concept of the AQG system from a laboratory grade text generation tool to an operationally self sustaining educational infrastructure component.

The three layer quality control pipeline, comprising algorithmic deduplication, expert human review, and drip publishing, likewise makes a substantive contribution to the practical AI in education literature. The hybrid human, AI validation model instantiated here reflects a theoretically grounded position: that current LLMs, whilst generatively powerful, remain susceptible to producing items requiring careful validation, and cannot yet be trusted to self validate the pedagogical suitability, distractor quality, and syllabus alignment of their outputs without human oversight, as the utility of LLM-based item generation relies heavily on a human in the loop to validate output and maintain pedagogical standards [27]. The platform's architecture neither overclaims full automation nor abdicates to wholly manual workflows; it occupies the theoretically and practically defensible middle ground of augmentation, a disposition that the educational technology literature increasingly endorses as the appropriate framing for AI deployment in high stakes contexts, as shown in Table 7 below:

Table 7. Summary of Key Findings and Their Primary Implications Across the Two Evaluative Dimensions

Dimension	Core Finding	Primary Implication
Student Performance Heterogeneity	CV of 17 until 28% across subtests; 20% of participants recorded floor scores in five of seven domains; composite average range: 350.00 until 834.43	Validates the necessity of adaptive, difficulty stratified question delivery; a single difficulty question bank cannot serve the competence spectrum present in the target population
Inter Subtest Score Asymmetry	Pronounced within student domain discrepancies (e.g., Radeikal: PK 832 vs LBE 504); all inter subtest r > 0.715, confirming a latent aptitude factor	Necessitates subtest specific diagnostic feedback; platform analytics should surface individual domain gaps rather than reporting composite averages alone
AQG Output Quality: Quantitative Domains	PK (96.3%), PM (92.9%), PPU (91.4%) acceptable generation rates; zero duplicate questions in all three; avg max similarity <42%	LLM based AQG is demonstrably viable for structured, schema constrained content; quantitative subtests represent the highest confidence deployment domain for automated pipeline operation
AQG Output Quality: Text Stimulus Domains	PBM (87.3%), LBI (80.4%), LBE (75.6%) duplication rates; acceptable rates of 9.5 until 17.8%; avg max similarity 84 until 90%	Jaccard based deduplication is insufficient for stimulus rich content; semantic similarity approaches (e.g., sentence transformer cosine distance) are required to distinguish surface lexical from substantive question overlap
Scheduler Driven Automation	1.663 questions generated without manual intervention; rotation logic maintained equitable subtest coverage; drip publishing sustained consistent daily output	Scheduler architecture is operationally feasible and cost manageable; represents a meaningful infrastructural advance over prompt on demand generation paradigms

3.3.5 Theoretical Implications: Advancing the Conceptual Foundations of AI Driven Educational Technology

Beyond its empirical contributions, this study advances the conceptual architecture of AI driven educational technology along several theoretically significant dimensions. First, and most fundamentally, it provides an empirically grounded instantiation of the personalised learning paradigm within a high stakes, nationally standardised examination context, a setting that has historically resisted personalisation owing to the uniformity imposed by centralised assessment design. The platform does not alter the examination itself; it adapts the preparatory environment to the learner's profile, thereby reconciling the structural tension between standardised

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

assessment and individualised pedagogy. This reconciliation is achieved not through learner facing curriculum redesign but through the back end mechanism of adaptive content generation and difficulty calibration, a modality that preserves the integrity of the standardised assessment whilst dramatically expanding the equity of access to high quality preparation resources.

Second, the study contributes to the theoretical discourse on AI content authenticity and psychometric validity in automated assessment. The difficulty calibration mechanism, which reclassifies question difficulty based on empirical accuracy rates rather than relying solely on a priori LLM assigned difficulty labels, represents a form of data driven psychometric anchoring that addresses a recognised weakness of purely generative approaches. LLMs, when left to self assign difficulty, tend to produce classifications that reflect surface textual complexity (lexical density, sentence length, numerical magnitude) rather than the cognitive demand actually experienced by learners. The empirical recalibration pipeline, operating on real learner response data, substitutes psychometric evidence for heuristic labelling, thereby strengthening the instrument's validity claims in a manner consistent with construct-oriented, psychometrically grounded evaluation frameworks that have been advocated as an alternative to purely task-based benchmarking of AI-generated content [28].

Third, this study contributes a theoretically principled critique of lexical similarity as a universal proxy for content uniqueness in educational assessment contexts. The differential performance of the Jaccard based deduplication algorithm across question types is not merely a practical engineering observation; it raises a deeper epistemological question about what 'uniqueness' means in the context of stimulus dependent comprehension items. A passage based reading question is unique not because its words differ from prior questions but because the cognitive operation it requires of the reader, the specific inferential move, evaluative judgement, or reflective act it demands differs. Defining uniqueness at the token level for such items is philosophically incoherent, and the platform's empirical data provide the evidentiary basis for formalising this critique. This finding invites the educational NLP research community to develop domain sensitive uniqueness taxonomies that distinguish between surface lexical, structural cognitive, and pedagogical intent dimensions of question distinctiveness.

3.3.6 Practical and Managerial Implications

3.3.6.1 Implications for Educational Technology Developers

For developers of AI driven educational platforms, the present study yields several design level prescriptions of immediate practical utility. The most urgent is the adoption of a domain differentiated deduplication strategy, replacing the current uniform Jaccard pipeline with a hybrid approach that applies word level Jaccard to quantitative and general knowledge subtests, where it performs demonstrably well, and semantic cosine similarity (or an ensemble of both) to passage dependent subtests. This modification alone would be expected to substantially recover the approximately 43% of flagged questions in text stimulus domains that are, in all likelihood, genuinely novel in their cognitive demand despite their surface lexical overlap. The implementation cost is modest relative to the likely yield in publishable content volume.

Developers should additionally consider the integration of a learner adaptive routing module at the platform front end, grounded in the competence stratification evidence documented in the present study. The data confirm that participants span an extraordinarily wide range of preparedness levels, from composite averages of 350.00 to 834.43 within a single cohort, and that a static difficulty distribution (30/40/30), however principled in design, cannot simultaneously serve the learner at floor and the learner approaching the ceiling. A diagnostic placement instrument, drawing upon the platform's existing accuracy rate data, could dynamically adjust each user's personal difficulty distribution in real time, approximating the Zone of Proximal Development targeting that theorists of intelligent tutoring have long proposed as the gold standard for adaptive learning systems.

At the infrastructure level, the scheduler architecture demonstrated in this study warrants adoption as a standard pattern for AQG based platforms intended to operate at scale. The four hourly generation cycle, combined with topic rotation and minimum stock logic, effectively solved the content freshness and distributional equity problems that beset on demand generation models without incurring prohibitive API costs. Developers operating under API quota constraints, an increasingly common operational reality for LLM dependent applications, as noted by Lim et al., will find the present system's batch size optimisation (five questions per execution cycle) a practically validated template for sustainable pipeline operation.

3.3.6.2 Implications for Educators and Academic Institutions

For educators and academic institutions engaged in UTBK-SNBT preparation, whether as bimbingan belajar (private tutoring) providers, school based preparation programmes, or university access offices, the empirical findings of this study carry implications that are simultaneously reassuring and cautionary. They are reassuring in their demonstration that AI generated questions in quantitative domains can achieve acceptable quality rates comparable to human authored content, suggesting that institutions with limited content development capacity can responsibly supplement their question banks with AQG output provided it passes through the platform's human review stage. The human review layer is not, in the present architecture, a redundant quality theatre; it is a structurally necessary intervention that catches the subset of AI generated questions that clear the deduplication

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

filter but nonetheless require pedagogical refinement, questions whose distractor logic is insufficiently discriminating, whose correct answers admit of ambiguity, or whose syllabus alignment is imprecise.

The findings are cautionary, however, in what they imply about the current limitations of full automation claims. Institutions that deploy AQG platforms without mandating expert human review, particularly for text stimulus content types, risk admitting into their practice banks questions that are technically nonduplicate but pedagogically deficient. The platform's architecture, in embedding human review as a nonnegotiable pipeline stage rather than an optional post hoc quality check, models the institutional governance standard that educational organisations should demand of any AI generated assessment content they deploy in high stakes preparation contexts. This is not a counsel of technophobia; it is a counsel of proportionate institutional due diligence.

Furthermore, the subtest specific performance profiles documented in this study provide educators with actionable intelligence regarding where AI assistance is most and least reliably deployable. The strong performance in PK, PM, and PPU indicates that AI generated question banks for these subtests can, with appropriate review, achieve the content volume and diversity required to sustain active recall based practice regimens, a pedagogically validated preparation strategy whose efficacy in improving examination outcomes is supported by the literature cited in Section 1 of this study. For PBM, LBI, and LBE, a more conservative deployment posture is warranted until semantic deduplication methodologies are implemented: human authored questions should continue to constitute the majority of the practice bank for these subtests, with AI generated content serving a supplementary rather than primary role.

3.3.6.3 Implications for Educational Policymakers

At the policy level, the findings of this study contribute to an evidence base that is increasingly urgently required to inform Indonesia's educational technology governance frameworks. The UTBK-SNBT's function as the primary gateway to state higher education, a gateway that, by design, admits fewer than one in three applicants, makes the equity dimensions of preparation resource access a matter of direct policy concern. The present study demonstrates that an AI driven preparation platform can, in principle, provide high quality, continuously refreshed, and diagnostically informed practice content at a marginal cost per learner that is structurally impossible to replicate through human authored content production. The implication for policy is that strategic investment in and quality certification frameworks for, such platforms has the potential to materially reduce the preparation resource disparities that currently correlate strongly with socioeconomic background among SNBT candidates.

Policymakers should, however, attend carefully to the quality assurance conditions under which such platforms operate. The present study's evidence that AQG performance varies substantially across subtest types, and that a naive interpretation of aggregate duplication rates may either overstate or understate system quality depending on the domain composition of the evaluation sample, counsels against the adoption of simple, single metric quality certification standards for AI generated educational content. A domain sensitive, multi dimensional quality framework, encompassing psychometric validity indicators (distractor quality, difficulty calibration accuracy, syllabus alignment verification) alongside the similarity based uniqueness metrics employed in the present study, would constitute a more epistemically robust basis for policy level certification of AQG platform deployments in the UTBK-SNBT preparation context.

3.3.6 Positioning the Present Study Against Prior Research

Recent studies on AI-assisted question generation have predominantly examined isolated capabilities rather than end-to-end, self-sustaining preparation ecosystems. Widjaja and Yohannis [10] demonstrated an AI-powered automatic question generation approach for teachers, yet their implementation remained dependent on manual prompt initiation at each generation cycle rather than autonomous, scheduled operation. Similarly, Henry [13], Elwirehardja and Pardamean examined automatic question generation for Bahasa Indonesia examinations using CopyNet, a study confined to a single language-generation architecture without an accompanying deduplication or scheduling framework. Al Faraby, Romadhony and Adiwijaya [15] established that large language models can generate items comparable in quality to those authored by human experts, but their evaluation was restricted to classification and generation accuracy rather than sustained, production-scale deployment. Lim, Gunasekara, Pallant, Pallant and Pechenkina [20] discussed the operational and resource implications of generative AI adoption in education from a managerial perspective, without proposing a concrete scheduling architecture to resolve such constraints. Ahmed et al. [22] reported that LLMs exhibit a tendency toward semantically repetitive output when uniqueness constraints are absent, a finding this study empirically substantiates through the differentiated Jaccard-based duplication rates observed across subtests. More recently, Kim and Park [24] surveyed the progression from syntactic to semantic text-deduplication techniques, underscoring the very limitation this study identifies in stimulus-rich subtests, while Crk and Gultepe [27] emphasised the continued necessity of human oversight in validating the instrumental quality of LLM-generated assessment items. Collectively, prior research has treated question generation, deduplication, scheduling, and human validation as separate concerns evaluated in isolation. The present study distinguishes itself by integrating these elements into a single, continuously operating platform evaluated across the complete structural diversity of the seven UTBK-SNBT subtests, thereby providing the first empirically grounded characterisation of how generation reliability varies systematically with question-type topology within one operational pipeline.

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

3.3.7 Limitations and Directions for Future Research

An intellectually honest appraisal of this study's contribution requires explicit acknowledgement of its limitations, each of which simultaneously defines a direction for future investigation. The most consequential limitation pertains to sample scope: the student performance data derive from a cohort of 30 participants, a scale that, whilst adequate for proof of concept characterisation, precludes generalisation to the broader UTBK-SNBT candidate population and does not permit the statistical power required for inferential testing of subgroup differences. Future studies should seek cohort sizes of several hundred participants, drawn from geographically and socioeconomically diverse backgrounds, to assess whether the performance variance patterns observed herein are stable characteristics of the population or artefacts of the present sample's specific composition.

A second limitation concerns the absence of a longitudinal component. The present study captures a single cross sectional snapshot of student performance without examining whether engagement with the platform over an extended preparation period produces measurable gains in subtest scores. The theoretical rationale for active recall based practice predicts exactly such gains, and the platform is architecturally equipped to collect the longitudinal usage and performance data required to test this prediction. A pre and post quasi experimental design, comparing SNBT outcomes between platform users and matched controls preparing through conventional methods, would constitute a substantially more consequential contribution to the efficacy literature than the present study's cross sectional analysis alone can achieve.

A third limitation, already addressed in Section 4.3, concerns the methodological adequacy of Jaccard based deduplication for stimulus rich content. Future work should implement and empirically compare semantic similarity approaches, particularly sentence transformer embedding cosine distance, against the present Jaccard baseline across all seven subtests, using expert human judgements of genuine question redundancy as the criterion variable. Such a comparison would provide the evidentiary foundation required to select and parameterise the optimal deduplication methodology for each subtest category, moving the platform's quality control architecture from its current heuristic configuration to an empirically grounded, domain differentiated design.

Finally, the AI Tutor and automated discussion components of the platform, both of which constitute potentially significant contributors to learning efficacy, have not been evaluated in the present study. Future investigations should develop and validate instruments for assessing the quality, pedagogical coherence, and learner perceived utility of AI Tutor interactions, situating these evaluations within established frameworks for conversational AI assessment in educational settings. The holistic evaluation of an integrated AI educational platform requires that each of its constituent components be subjected to rigorous individual and systemic scrutiny, a research agenda that the present study initiates rather than exhausts.

4. CONCLUSION

This study demonstrates that the proposed scheduler driven platform can sustain continuous, adaptive question generation for the UTBK-SNBT, while also revealing where this capability currently ends. The heterogeneity observed across the 30 participant cohort, including the presence of floor scores among a fifth of participants, indicates that the platform's adaptive design addresses genuine variation in learner readiness, though the small and non randomly sampled cohort restricts these observations to descriptive characterisation rather than generalisable evidence of learning impact. The disparity in AQG performance across subtests is more consequential than a simple metric gap: for schema constrained domains such as Quantitative Knowledge and Mathematical Reasoning, Jaccard based deduplication approximates true uniqueness reasonably well, but for stimulus rich, reading intensive domains such as Literacy in Indonesian, Literacy in English, and Reading Comprehension and Writing, the same metric systematically overstates duplication because it cannot recognise paraphrased or semantically equivalent passages. In practice, this means the platform cannot yet generate reliably unique long text stimuli without additional human review, a limitation that directly constrains its readiness for unattended deployment in text heavy subtests. A further limitation concerns the LLM itself: its training data and reasoning patterns are not specifically calibrated to the structure, register, and content emphases of the Indonesian SNBT curriculum, so generated items in culturally or curriculum specific contexts may require closer alignment checks than this study was able to perform. Taken together, the platform should be regarded as a feasible foundation for further development rather than a validated, scalable solution, with semantic similarity based deduplication, expert pedagogical review, and curriculum specific evaluation identified as necessary next steps before broader deployment can be responsibly claimed.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to all Sobat TELISIK, Tutors, Parents/Guardians, Partners, and Colleagues who have continuously supported the implementation of this research. Special appreciation is extended to Mrs. Amaliah Sobli, S.KG., M.B.A., as a benefactor of TELISIK, for her invaluable assistance, unwavering support, and genuine concern for the advancement of the TELISIK educational and literacy community in South

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 979–995
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10356

Sumatra. Her encouragement and commitment have made a meaningful contribution to the successful completion of this research. Thank you to everyone who has supported the implementation of this research.

REFERENCES

[1] Muhammad Syauqi Firdaus, Wahyu Kholis Prihantoro, Latifaturrohmah Latifaturrohmah, Husna Rifaatul Mahmudah, and Afrida Aunil Illah, "Efektivitas Instrumen Tes Uraian Dibandingkan dengan Tes Pilihan Ganda dalam Mengukur Hasil Belajar Siswa di MAN 2 Bantul," Jurnal Kajian Ilmu Pendidikan, Bahasa dan Komunikasi, vol. 1, no. 4, pp. 87–101, Nov. 2025, doi: 10.61132/jkaipbaku.v1i4.175.
[2] S. Zhang, Z. Wang, J. Qi, J. Liu, and Z. Ying, "Accurate Assessment via Process Data," Psychometrika, vol. 88, no. 1, pp. 76–97, Mar. 2023, doi: 10.1007/s11336-022-09880-8.
[3] M. A. N. Awal, Hasniati, Husni Agriani, and Izmi Alwiah M, "Pelatihan Mengerjakan Tes Potensi Skolastik Bagi Siswa Kelas XII SMA Untuk Menghadapi Seleksi Nasional Berdasarkan Tes (SNBT)," Journal Inovasi Pengabdian Masyarakat, vol. 1, no. 3, pp. 126–132, Oct. 2024, doi: 10.65255/jipmas.v1i3.102.
[4] Seleksi Nasional Penerimaan Mahasiswa Baru, "Informasi Umum | Seleksi Nasional Penerimaan Mahasiswa Baru." Accessed: May 15, 2026. [Online]. Available: https://snpmb.id/utbk-snbt/informasi-umum
[5] Seleksi Nasional Penerimaan Mahasiswa Baru, "Statistik Nilai Peserta UTBK 2024," Jun. 2024. Accessed: Jun. 15, 2026. [Online]. Available: https://snpmb.id/blog/statistik-nilai-peserta-utbk-2024
[6] Seleksi Nasional Penerimaan Mahasiswa Baru (SNPMB), "SIARAN PERS Nomor:07/sipers/snpmb/V/2025 TENTANG HASIL SNBT 2025 RESMI DIUMUMKAN: 29,43% PESERTA DINYATAKAN LULUS ," May 2025. [Online]. Available: https://files.snpmb.id/web2025/07_Sipers%20Pengumuman%20SNBT%202025.pdf
[7] J. Xu et al., "Active recall strategies associated with academic achievement in young adults: A systematic review," J. Affect. Disord., vol. 354, pp. 191–198, Jun. 2024, doi: 10.1016/j.jad.2024.03.010.
[8] J. Schwerter, F. Lauermann, T. Brahm, and K. Murayama, "Differential use and effectiveness of practice testing: Who benefits and who engages?," Learn. Individ. Differ., vol. 123, p. 102761, Oct. 2025, doi: 10.1016/j.lindif.2025.102761.
[9] J. Burstein, R. Cardwell, P.-L. Chuang, A. Michalowski, and S. Nydick, "Exploring AI-Enabled Test Practice, Affect, and Test Outcomes in Language Assessment," ArXiv, Aug. 2025.
[10] S. Y. Widjaja and A. Yohannis, "AI-Powered Automatic Question Generation for Teachers," in 2025 International Conference on Smart Computing, IoT and Machine Learning (SIML), IEEE, Jun. 2025, pp. 1–6. doi: 10.1109/SIML65326.2025.11081160.
[11] B. Das, M. Majumder, S. Phadikar, and A. A. Sekh, "Automatic question generation and answer assessment: a survey," Res. Pract. Technol. Enhanc. Learn., vol. 16, no. 1, p. 5, Dec. 2021, doi: 10.1186/s41039-021-00151-1.
[12] P. Panchal, J. Thakkar, V. Pillai, and S. Patil, "Automatic Question Generation and Evaluation," Journal of University of Shanghai for Science and Technology, vol. 23, no. 05, pp. 751–761, May 2021, doi: 10.51201/JUSST/21/05203.
[13] M. M. Henry, G. N. Elwirehardja, and B. Pardamean, "Automatic question generation for bahasa indonesia examination using copynet," Procedia Comput. Sci., vol. 245, pp. 953–962, 2024, doi: 10.1016/j.procs.2024.10.323.
[14] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," Int. J. Artif. Intell. Educ., vol. 30, no. 1, pp. 121–204, Mar. 2020, doi: 10.1007/s40593-019-00186-y.
[15] S. Al Faraby, A. Romadhony, and Adiwijaya, "Analysis of LLMs for educational question classification and generation," Computers and Education: Artificial Intelligence, vol. 7, p. 100298, Dec. 2024, doi: 10.1016/j.caeai.2024.100298.
[16] Seleksi Nasional Penerimaan Mahasiswa Baru (SNPMB), "Ujian Tertulis Berbasis Komputer (UTBK) dalam Seleksi Nasional Berdasarkan Tes (SNBT) Tahun 2026."
[17] L. Reynolds and K. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," ArXiv, 2021.
[18] PISA 2022 Assessment and Analytical Framework. OECD Publishing, 2023. doi: 10.1787/dfe0bf9c-en.
[19] B. S. Bloom, Taxonomy of educational objectives. 2023.
[20] W. M. Lim, A. Gunasekara, J. L. Pallant, J. I. Pallant, and E. Pechenkina, "Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators," The International Journal of Management Education, vol. 21, no. 2, p. 100790, Jul. 2023, doi: 10.1016/j.ijme.2023.100790.
[21] M. N. Salem and M. F. Abdullah, "Evaluating String-Based Similarity Algorithms for Duplicate Record Identification in Databases," Journal of Science and Technology, vol. 31, no. 2, Feb. 2026, doi: 10.20428/jst.v31i2.3566.
[22] W. M. Ahmed, A. A. Azhari, A. Alfaraj, A. Alhamadani, M. Zhang, and C.-T. Lu, "The Quality of AI-Generated Dental Caries Multiple Choice Questions: A Comparative Analysis of ChatGPT and Google Bard Language Models," Heliyon, vol. 10, no. 7, p. e28198, Apr. 2024, doi: 10.1016/j.heliyon.2024.e28198.
[23] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du, "Artificial intelligence in education: A systematic literature review," Expert Syst. Appl., vol. 252, p. 124167, Oct. 2024, doi: 10.1016/j.eswa.2024.124167.
[24] J. Kim and Y. Park, "A Survey of Text Deduplication: From Syntactic Matching to Semantic Understanding," IEEE Access, vol. 14, pp. 16731–16750, 2026, doi: 10.1109/ACCESS.2026.3658439.
[25] 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, p. 473, Mar. 2024, doi: 10.62527/joiv.8.1.2217.
[26] J. Risch, T. Möller, J. Gutsch, and M. Pietsch, "Semantic Answer Similarity for Evaluating Question Answering Models," Oct. 2021.
[27] I. Crk and E. Gultepe, "Evaluating the instrumental quality of LLM-generated assessment items," Front. Educ. (Lausanne)., vol. 11, Jun. 2026, doi: 10.3389/feduc.2026.1837523.
[28] X. Wang et al., "Evaluating General-Purpose AI with Psychometrics," Dec. 2023.