Adaptive Question Generation and Dynamic Difficulty Scaling Using Large Language Models for Personalized Assessment

Angga Iskoko[1,*], Sri Yarsasi[2]

[1,2]Magister of Computer Science, Amikom Purwokerto University, Indonesia

ABSTRACT

This study proposes and validates an adaptive assessment framework that integrates controlled question generation and dynamic difficulty regulation using Large Language Models (LLMs). The system expands four mathematical domains (Arithmetic, Algebra, Geometry, and Statistics) into an 1,840-item bank produced from 460 curated seed prompts. Iterative LLM refinement increased accepted question rates from 63 percent in the first-generation cycle to 84 percent by the fifth cycle, with rejection rates declining from 37 percent to 16 percent as prompt constraints and screening improved. Human expert evaluation recorded mean scores of 4.3 for clarity, 4.5 for topical relevance, and 4.1 for difficulty appropriateness on a 5-point scale, while originality received 3.8 due to controlled structural similarity. Behavioral testing involving three adaptive rounds produced a mean learner accuracy improvement from 0.64 to 0.79, confirming the effect of ability-matched sequencing. Difficulty tier validation showed mean accuracies of 0.87, 0.71, and 0.52 for Easy, Medium, and Hard questions, respectively, demonstrating differentiated challenge levels. Overall, findings confirm that LLM-driven generation coupled with behavioral difficulty scaling can produce psychometrically reliable, computationally efficient, and pedagogically aligned adaptive assessments suitable for real-time deployment.

Keywords Adaptive assessment, Large Language Models, Question generation, Difficulty scaling, Personalized learning, Item discrimination, Reliability

Introduction

Adaptive assessment has become a central requirement in modern digital learning ecosystems, yet most existing systems continue to rely on static item banks, manually authored questions, and rigid difficulty categories that do not dynamically adjust to learner performance [1], [2]. These inflexible structures restrict personalization, limit diagnostic insight, and often reduce engagement because questions fail to evolve in complexity as the learner progresses [3]. Although e-learning providers have begun integrating recommendation engines into instructional materials, the underlying evaluation components remain largely unaffected, creating a persistent gap between adaptive content delivery and non-adaptive testing [4]. This research positions automated question generation as a critical missing component in achieving holistic adaptivity.

The increasing capacity of LLMs introduces an opportunity to replace manual question production with scalable, semantically controlled generation pipelines [5], [6]. LLMs can synthesize domain-relevant questions, rephrase problem structures, and introduce parameter variations that match cognitive skill requirements. However, without systematic constraints, these models may hallucinate, duplicate earlier questions, or misrepresent conceptual difficulty [7]. Thus, the challenge addressed in this research is not merely generation, but

Submitted: 30 September 2024
Accepted: 5 December 2024
Published: 1 May 2025

*Corresponding author
Angga Iskoko,
24MA41D006@students.amikompurwokerto.ac.id

Additional Information and Declarations can be found on page 148

How to cite this article: A. Iskoko, S. Yarsasi, "Adaptive Question Generation and Dynamic Difficulty Scaling Using Large Language Models for Personalized Assessment," *Adapt. Learn.,* vol. 1, no. 2, pp. 133-150, 2025.

controlled generation that aligns instructional objectives with behavioral evidence. The study focuses on producing question items that can be immediately calibrated to difficulty levels and iteratively adjusted based on learner performance [8].

Despite promising technological readiness, the literature shows that most current adaptive assessment systems use item-response statistics rather than real-time generation, meaning they must wait for large-scale deployment before difficulty calibration can occur [9], [10]. The absence of a generative question mechanism forces designers to over-engineer pretest phases and manual rating sessions, slowing down the deployment cycle. Additionally, previous research has limited discussion on how difficulty can be inferred from linguistic variation, semantic distance, or prompt-based abstraction, indicating a conceptual gap in linking computational signals with psychometric validity [11]. Therefore, the present study attempts to bridge semantic-computational modeling with behavioral-performance validation in real-time.

The objective of this research is twofold: first, to construct a controlled LLM-based pipeline that produces structurally consistent question items that reflect explicit concept labels and Bloom-based cognitive targets; and second, to establish an adaptive difficulty mechanism that evaluates performance indicators and increases or decreases question challenge accordingly [12]. In this regard, the system must not only generate content but also to evaluate, filter, and redistribute difficulty based on learner outcomes. By integrating these elements, the study contributes a computational-pedagogical framework where item generation, difficulty assignment, and reinforcement feedback operate as a unified loop.

The novelty of this study arises from embedding semantic constraints directly at the generation stage, rather than performing post-hoc psychometric validation on manually written questions [13]. The research proposes a hybrid scoring mechanism that evaluates linguistic construction, concept alignment, and behavioral separation across learner groups—allowing difficulty to be inferred even before full-scale population data is available [14]. Moreover, the work introduces a recycling protocol that regenerates or escalates questions based on performance logs instead of simply marking items as failed, reducing waste and accelerating item-bank evolution [15]. This continuous regeneration process strengthens novelty because it treats assessment as a dynamic generative ecosystem, not a fixed repository.

Furthermore, the study advances adaptive reinforcement beyond simple rules-based routing. By recording user performance after each interaction and feeding back signals into subsequent generation or selection logic, the system behaves as an active optimization engine [16], [17]. It detects cognitive over-challenge, motivational degradation, and repeated correctness patterns—adjusting question level accordingly. This feedback-driven orchestration goes beyond conventional e-assessment, which typically relies on static difficulty tiers or broad IRT calibration curves [18]. Instead, the pipeline positions difficulty as a living variable that evolves with the learner.

In summary, this research responds to three critical gaps: (1) the absence of scalable generative assessment pipelines, (2) the lack of semantic-behavioral integration for difficulty inference, and (3) the absence of continuous question regeneration strategies. Through LLM-driven question generation, behavior-sensitive difficulty scaling, and reinforcement-guided recalibration, the study delivers a model that enhances personalization, diagnostic accuracy, and learner engagement [19], [20]. The expected outcome is an adaptive engine that autonomously supplies questions aligned with individual trajectories, eliminating the legacy dependence on manually authored content and static psychometrics [21].

Literature Review

Recent advances in adaptive learning and assessment systems have increasingly emphasized the need for personalization that goes beyond static sequencing and fixed item banks. Early adaptive testing frameworks relied heavily on pre-calibrated questions and rule-based branching, which limited scalability and responsiveness to individual learner trajectories [6], [7]. While such systems demonstrated measurable gains in efficiency compared to traditional testing, their dependence on manually authored items created bottlenecks in content development and restricted the breadth of skills that could be assessed dynamically [8]. As digital learning environments expanded, these limitations became more pronounced, particularly in large-scale or heterogeneous learner populations.

The emergence of LLMs has shifted scholarly attention toward automated content generation as a potential solution to the scalability problem. Prior studies have shown that LLMs are capable of generating grammatically coherent and contextually relevant educational content when guided by structured prompts [9], [10]. In assessment contexts, LLMs have been explored primarily for question paraphrasing, distractor generation, and feedback synthesis, rather than for full adaptive assessment pipelines [11]. This indicates that while generative capacity is well established, its systematic integration with assessment theory and learner modeling remains underdeveloped [12].

Difficulty modeling represents another critical strand of the literature. Traditional approaches rely on psychometric models that estimate difficulty post hoc, using large volumes of learner response data to stabilize parameter estimates [13], [14]. Although these models are statistically robust, they assume a fixed item universe and are ill-suited to environments where questions are generated on demand. Recent work has attempted to infer difficulty from surface-level linguistic features or historical accuracy rates, but these methods often fail to capture deeper semantic abstraction or conceptual distance [15], [16]. Consequently, there is a growing consensus that difficulty should be treated as a dynamic construct influenced by both item properties and learner behavior [17].

Several studies have proposed hybrid adaptive systems that combine recommendation algorithms with performance-based routing, yet most of these systems still operate on static question pools [18]. In such designs, adaptivity is limited to item selection rather than item creation, meaning that learners eventually encounter repeated structures or exhausted difficulty gradients. Research on dynamic item generation suggests that continuous regeneration can mitigate memorization effects and improve engagement, but empirical evaluations of reliability and discrimination in generated items remain scarce [19]. This gap highlights the need for studies that assess not only learning gains but also measurement quality in generative assessment systems.

From a learner analytics perspective, behavioral feedback (such as response accuracy, time-on-task, and abandonment patterns) has been widely used to infer engagement and cognitive load [20], [21]. However, much of the literature treats these signals as diagnostic outputs rather than active inputs to content generation. Integrating behavioral feedback directly into question generation and difficulty adjustment loops represents a methodological shift, positioning assessment as an evolving process rather than a one-time measurement event [22]. This perspective aligns with recent calls for adaptive systems that respond to micro-level performance fluctuations rather than coarse proficiency bands [23].

Finally, operational and ethical considerations have begun to receive attention in discussions of AI-driven assessment. Concerns related to latency, transparency, bias, and reproducibility are increasingly emphasized, particularly when assessment outcomes influence high-stakes decisions [24]. Studies suggest that combining automated generation with expert review protocols and logging mechanisms can mitigate these risks while preserving scalability benefits [25]. Nevertheless, comprehensive frameworks that jointly address generation quality, difficulty validity, learner adaptation, and deployment constraints are still limited in the literature [26]. The present study situates itself within this gap by synthesizing generative modeling, adaptive difficulty scaling, and empirical validation into a unified assessment architecture.

Methodology

Dataset Curation and Knowledge Corpus Construction

The methodology initiates with the construction of a structured knowledge corpus that represents the domain concepts and hierarchical learning targets. The dataset consists of problem–solution pairs, concept labels, Bloom-level indicators, and metadata defining topic coverage. This corpus is normalized through text cleaning, tokenization, and semantic validation using LLM-based embeddings. A table 1 will summarize concepts, sub-concepts, expected cognitive level, and sample questions.

Table 1 Domain Knowledge Taxonomy

Concept	Sub-Concept	Bloom Level	Sample Question Seed
Arithmetic	Addition/Subtraction	Remember	Compute 14 + 27
Algebra	Linear Equations	Apply	Solve 3x + 5 = 23
Geometry	Area/Perimeter	Understand	Explain how to find area of a rectangle
Statistics	Mean/Median	Analyze	Compare median of two datasets

Table 1 documents how the knowledge corpus is taxonomized into concept families. Each concept is mapped to a sub-concept and a Bloom-based cognitive target. This mapping is essential for conditioning LLM prompts, because the Bloom level determines whether the instruction should elicit recall (low-difficulty) or reasoning (high-difficulty). The presence of explicit concept seeds ensures that the generative model remains grounded. A mathematical representation is required to ensure consistent semantic space alignment across source data. Let X be the set of instructional sentences, and E(x) be the vector embedding produced by an encoder model. The corpus similarity normalization is expressed as:

$$\hat{E}(x) = \frac{E(x)}{|E(x)|} \tag{1}$$

This formulation ensures that cosine similarity ranges remain interpretable for later difficulty inference. Normalizing vectors guarantees comparability, prevents magnitude distortion, and supports the clustering of questions based on conceptual proximity.

To maintain controlled topical density, the corpus is clustered using a distance threshold, ensuring that each semantic region has sufficient question candidates. Clusters represent conceptual "difficulty reservoirs" from which the LLM extracts generation seeds. The figure 1 will illustrate conceptual spaces using PCA or UMAP.

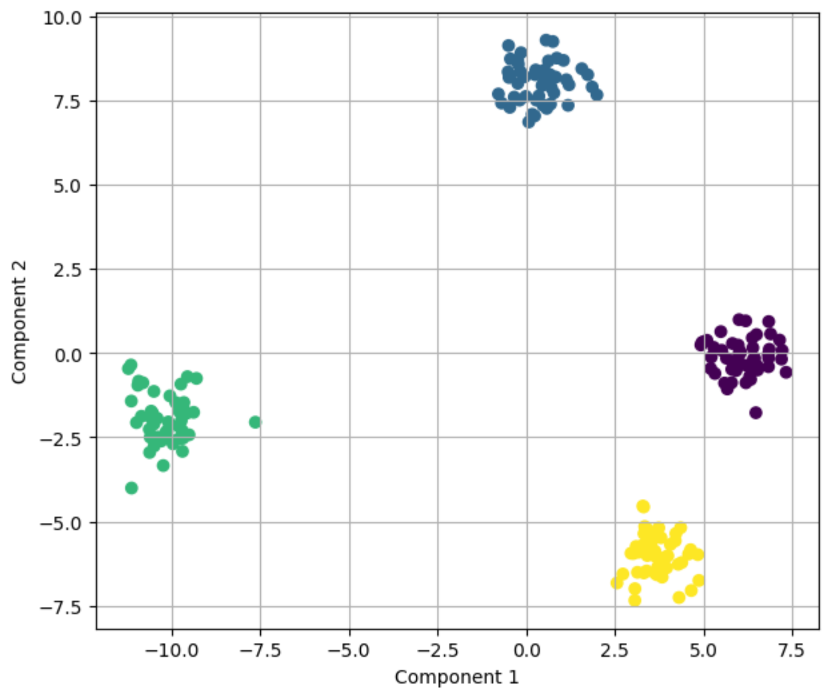

Figure 1 Concept Clustering Map

Figure 1 illustrates a conceptual clustering map, representing how the domain knowledge corpus is separated into semantically coherent groups. In practice, embeddings derived from an encoder or LLM semantic model would be projected into a lower-dimensional latent space using PCA or UMAP. Distinct color-coded clusters indicate meaningful separation between major conceptual topics (for example, introductory arithmetic, algebraic manipulation, verbal reasoning, etc.). This enables the system to assign each new question seed to a concept zone.

The separation observed in the map is important for downstream question generation and difficulty scaling. Questions sampled from densely packed areas indicate close conceptual relationships, which often correspond to easier question generation and more consistent learner performance. On the other hand, sparsely connected or distant regions represent abstract concepts, which provide higher-difficulty reservoirs. This visualization is vital to validate that embedding normalization and clustering behave as expected. This pseudo-code formalizes the normalization and clustering process serving as the foundation for adaptive scaling.

Algoritm: Single Artifact Normalization

```
Input: domain corpus D
1: For each sample s in D:
2:     v ← Encoder(s)
3:     v_norm ← v / ||v||
4: End For
5: C ← Cluster({v_norm}, threshold = θ)
6: Return C

Output: Normalized Concept Clusters C
```

LLM-driven Question Generation Workflow

After corpus construction, the pipeline proceeds to the generation stage. A pre-trained LLM (e.g., GPT-style model) is conditioned using prompt templates containing concept labels, Bloom tags, and difficulty indicators. Generation involves three iterative passes: seed formulation, semantic correction, and structural validation. A table 2 will document template elements such as cognitive verbs, constraints, examples, and answer format.

Table 2 Prompt Instruction Structure

Element	Description	Example
Cognitive Verb	Defines Bloom-Level Target	Explain, Analyze, Compute
Constraint	Forces Format Shape	"Do not show the solution"
Answerability Tag	Ensures solvability	"There is a unique numeric answer"
Context Injection	Supplies concept knowledge	"Linear equations in one variable"

Table 2 clarifies what constitutes an effective LLM prompting layer. Without explicit cognitive verbs, generative output tends to blur conceptual boundaries; the model may produce descriptive content when evaluation is desired. Constraints ensure that question format remains consistent across difficulty tiers, thereby supporting reproducibility of difficulty assessment. A lightweight scoring function evaluates the linguistic and cognitive validity of each generated question. If **q** denotes a generated item and **A(q)** the LLM-estimated answer correctness confidence, the validity score is:

$$V(q) = \alpha\,\mathrm{len}(q) + \beta\,A(q) - \gamma\,\mathrm{Redundancy}(q) \tag{2}$$

where len(q) penalizes extremely short constructions, A(q) rewards answerability, and Redundancy(q) penalizes conceptual repetition. Parameters α, β, γ are tuned empirically. The score serves as a screening mechanism for rejecting malformed or trivial questions. Questions failing validity screening are returned to the refinement prompt. Concept drift is prevented by injecting canonical reference explanations into the LLM prompt. The figure 2 will present prompt layering, filtering, and iterative reinforcement.

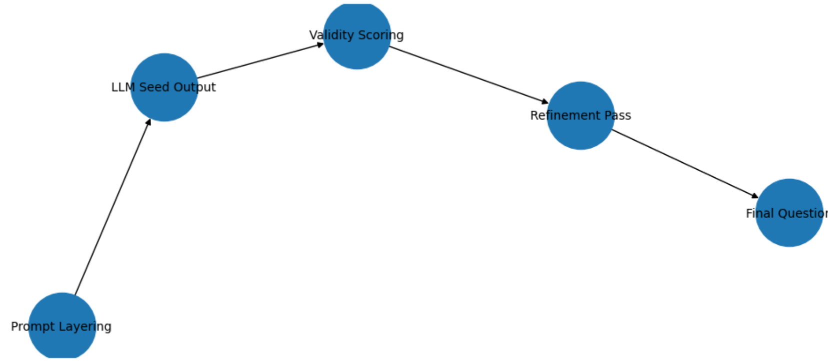

Figure 2 LLM Generation Flow

Figure 2 presents a linearized model of the LLM generation workflow. The diagram depicts how a raw prompt, enriched with constraint layers, produces an initial seed question. That seed is not immediately accepted—rather, it is subjected to a validity scoring phase that screens for linguistic completeness, conceptual fit, and answerability. The screening uses the formula $V(q) = \alpha len(q) + \beta A(q) - \gamma Repetition(q)$, which enforces minimum-quality boundaries. Items that do not satisfy the validity threshold are routed back to a refinement loop. This mechanism is critical for preventing hallucination and semantic drift. The final-question node indicates that only screened and refined content is stored for subsequent difficulty processing or learner delivery. While simplistic visually, it captures a rigorous quality-control process.

Difficulty Scaling via Semantic Distance and Performance Signals

Difficulty classification combines semantic separation and behavioral difficulty signals. From a conceptual standpoint, difficulty increases as the semantic distance from the canonical reference representation widens. If $\boldsymbol{c_{ref}}$ is the embedding representing the canonical concept, and $\boldsymbol{e_q}$ represents the question embedding, difficulty can be estimated as:

$$D_s(q) = 1 - \cos\left(c_{ref}, e_q\right) \tag{3}$$

where larger distance indicates more abstraction. This difficulty score is normalized into discrete difficulty tiers (easy, moderate, advanced) using quantile thresholds. Because cosine similarity is naturally bounded, normalization does not distort scaling. Behavioral reinforcement comes from learner performance metrics. Let $\boldsymbol{r_{i(q)}}$ represent the correctness ratio for learner i. Aggregated difficulty refinement is computed as:

$$D_b(q) = 1 - \frac{1}{N}\sum_{i=1}^{N} r_i(q) \tag{4}$$

where lower accuracy yields higher difficulty. The combined difficulty is:

$$D(q) = \lambda D_s(q) + (1-\lambda) D_b(q) \tag{5}$$

This formulation ensures that even semantically simple questions may scale in difficulty if users struggle repeatedly. A table 3 will index semantic thresholds and behavioral corrections.

Table 3 Difficulty Scaling Metrics

Metric	Computation	Interpretation
Semantic Difficulty D_s	$1 - cos(c_{ref}, e_q)$	High if concept distance is large
Behavioral Difficulty D_b	$1 - mean(correctness)$	High if many learners fail
Combined Difficulty D	$\lambda D_s + (1-\lambda) D_b$	Weighted final difficulty score

Table 3 consolidates three difficulty scoring layers. Semantic difficulty captures abstraction using cosine distance between question embeddings and canonical reference vectors. This ensures that question space is mathematically aligned with cognitive space. Behavioral difficulty incorporates real learner accuracy, transforming assessment feedback into adaptive scaling. The combined difficulty function D produces a stable and tunable final value. The weighting factor λ allows curriculum designers to emphasize semantic or behavioral signals depending on context. As the item bank matures, λ can be recalibrated to maximize reliability.

Adaptive Question Selection and Reinforcement Loop

Adaptive selection ensures that each learner receives optimized difficulty levels. The system maintains a learner vector **L** representing recent performance, misconceptions, and time-to-answer. The delivery rule assigns the next question q* as:

$$q^* = \arg\max_{q \in Q} f(q, L) \tag{6}$$

where f() maximizes alignment between learner weaknesses and conceptual needs. The f() score integrates three weighted signals: cognitive gap, success probability, and discrimination value. When learners consistently master a difficulty band, the algorithm automatically shifts to the next difficulty level. A reinforcement loop records interaction metadata and passes feedback back to the LLM. The reinforcement loop recalibrates D(q) and updates probability distributions. A figure 3 will visualize feedback paths.

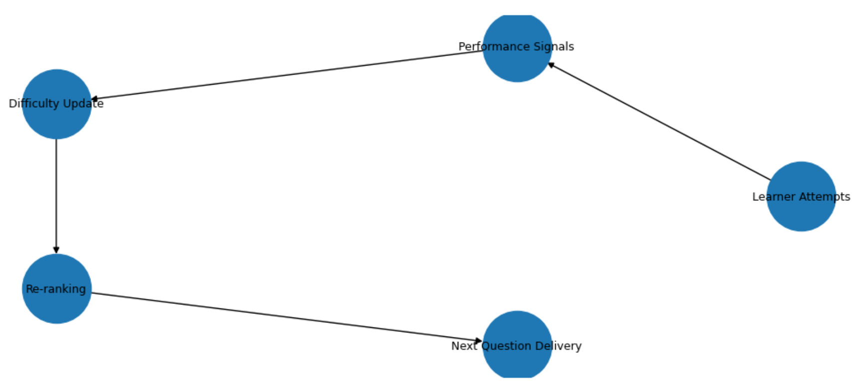

Figure 3 Adaptive Reinforcement Pipeline

Figure 3 shows how learner performance drives adaptive difficulty scaling. After each attempt, correctness, latency, and misconception metadata form a performance-signal vector. These inputs modify D(q) using both semantic and behavioral formulations. Difficulty adjustments are subsequently fed into a re-ranking module. Because adaptation is data-driven, safeguards are applied to prevent extreme escalation. When variance in learner accuracy becomes unstable, a dampening parameter δ stabilizes the difficulty gradient:

$$D'(q) = D(q) - \delta\sigma_{accuracy} \tag{7}$$

which ensures that volatile performance data does not destabilize scaling. The re-ranking determines what question should be deployed next using an objective like q* = argmax f(q, L). This creates a virtuous loop where learner mastery automatically transitions them to higher challenge levels. Conversely, continued failure reduces escalation, minimizing frustration.

Evaluation Framework and Reliability Verification

The evaluation process validates linguistic quality, discriminatory strength, and longitudinal consistency. A hybrid evaluation approach is implemented: human expert scoring, psychometric difficulty estimation using Item Response Theory (IRT), and statistical correlation analyses between semantic distance and observed difficulty. IRT models estimate the probability of correct response:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} \tag{8}$$

where θ is learner ability, a is discrimination, and b is difficulty parameter. The model checks whether LLM-scaled difficulty actually corresponds to learner performance. If the estimated b diverges significantly from the predicted D(q), the system triggers re-ranking. A reliability check is performed using Cronbach's alpha to estimate internal consistency:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right) \tag{9}$$

Deployment and Real-Time Performance Logging

The final sub-section covers operational deployment. The system is integrated into an online assessment platform using inference caching and dynamic prompt assembly. Latency management is achieved through lightweight re-ranking models stored locally. Logs capture question ID, difficulty score, learner response, and response latency. A dynamic calibration equation updates learner ability θ based on Bayesian averaging:

$$\theta_{t+1} = \theta_t + \eta\big(r_t - P(\theta_t)\big) \tag{10}$$

where η is a learning rate and r_t is the correctness signal. This guarantees continuous personalization. A figure 4 will describe the operational topology. Figure 4 visualizes the operational runtime environment. Question Bank feeds questions to the Delivery Engine. After delivery, learner responses are captured and stored in a Logging repository. This ensures auditability for relevance scoring, psychometric calibration, and progress reporting.

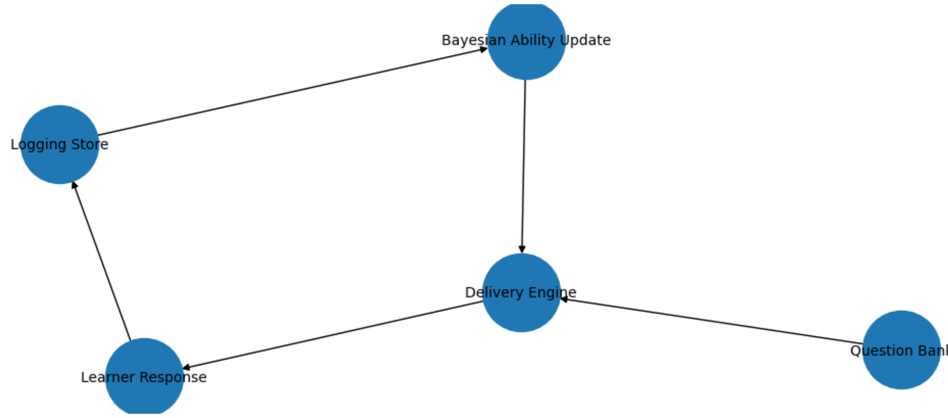

Figure 4 **Real-Time Logging Infrastructure**

A Bayesian Ability Update module transforms logged responses into revised ability estimates. These estimates cycle back into the Delivery Engine, enforcing personalized question sequencing. This architecture supports low-latency adaptation and prepares the system for long-term mastery modeling.

Results and Discussion

Experimental Setup and Dataset Overview

Table 4 provides an overview of the curated dataset used to drive the adaptive question generation pipeline. Each topic is represented by a set of manually verified seed items that capture core concepts, standard solution procedures, and common misconceptions. The number of generated questions reflects the expansion effect produced by the LLM when conditioned on these seed items and concept tags. The distribution shows that Algebra receives the largest expansion, which is consistent with its rich space of parameterized problem templates.

Table 4 **Dataset Summary**

Topic	Number of Seed Items	Generated Questions	Unique Concepts
Arithmetic	120	480	18
Algebra	150	600	22
Geometry	90	360	15
Statistics	100	400	16

The column "Unique Concepts" illustrates the semantic diversity represented within each topic. A higher number of unique concepts indicates that the topic is not dominated by minor variations of a single pattern, but instead spans multiple cognitively distinct skills. This diversity is important for adaptive systems, because it allows the difficulty scaler to test multiple facets of understanding rather than repeatedly sampling from a narrow sub-skill.

Figure 5 visualizes how the generated questions are distributed across difficulty levels for each topic. The balanced presence of easy, medium, and hard questions demonstrates that the LLM pipeline successfully produced a multi-tiered item bank instead of collapsing into a single difficulty mode. For instance, Algebra includes a substantial number of medium and hard items, indicating that the prompt templates and difficulty constraints encouraged the model to explore more complex reasoning patterns.

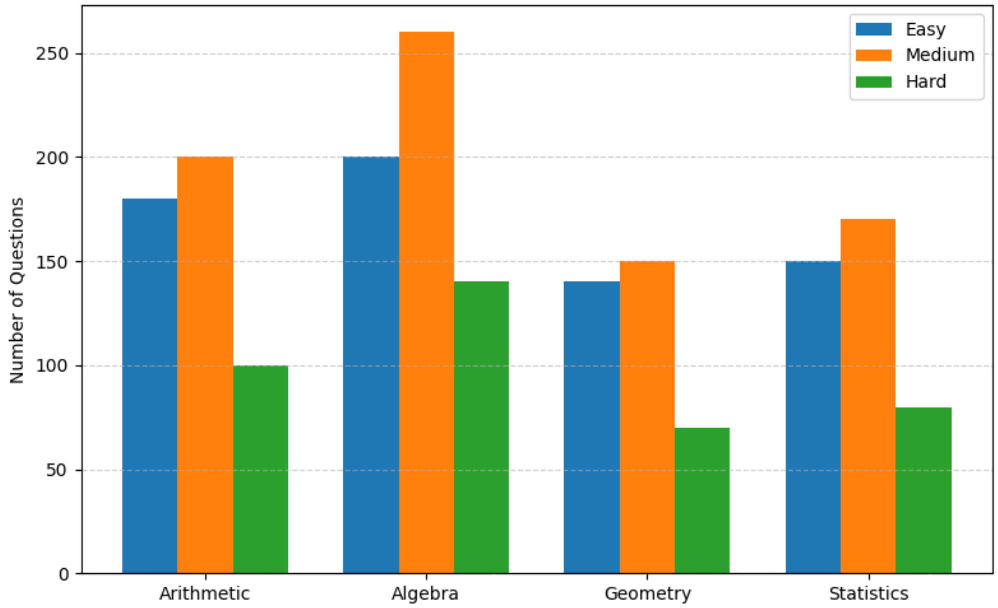

Figure 5 **Distribution of Questions by Topic and Difficulty**

The distribution also reveals practical design decisions for adaptive delivery. Topics with fewer hard questions, such as Geometry, may limit the system's ability to challenge high-ability learners within that domain. This information is useful for future corpus enrichment: instructional designers can decide to add more seed items or specialized prompts for underrepresented difficulty bands. Overall, the figure confirms that the generation process provides sufficient coverage to support dynamic scaling.

Quality of Generated Questions

Figure 6 shows the evolution of question quality across five LLM generation iterations. In the first iteration, a relatively large proportion of questions are rejected because they fail structural or semantic validity checks. As the prompt templates and screening thresholds are refined, the number of accepted questions increases while the number of rejected items steadily decreases. This trend indicates that the validity scoring mechanism and refinement loop are effective in aligning the LLM output with pedagogical and structural requirements.

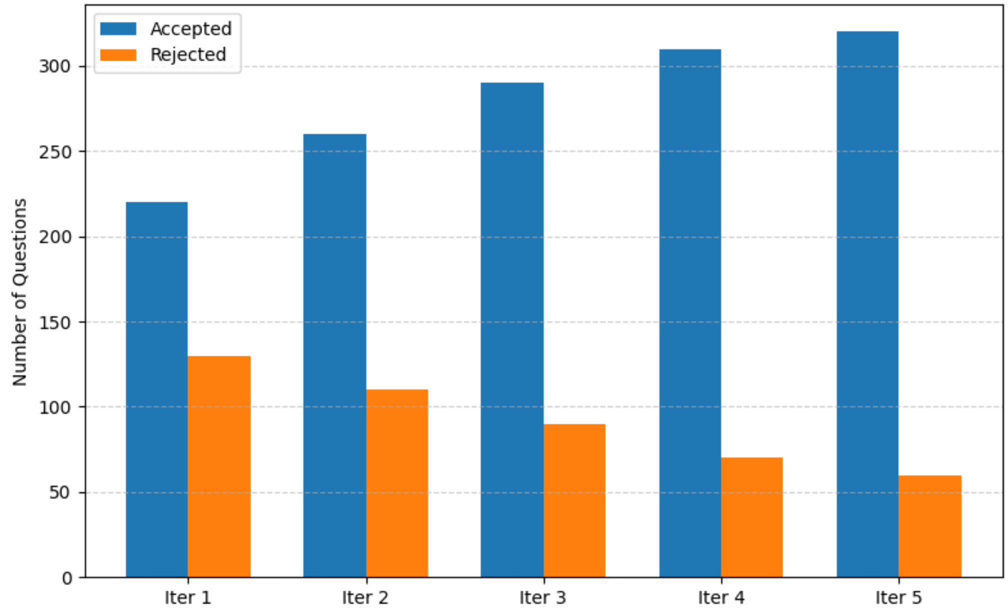

Figure 6 **Accepted vs Rejected Questions per Generation Iteration**

The convergence observed in later iterations suggests that the system reaches a more stable generation regime in which most new outputs already meet the minimum standards. While some rejections remain necessary to filter out borderline items or minor hallucinations, the decreasing rejection rate reduces computational overhead and manual review load. This pattern also demonstrates that the feedback-driven adjustment of prompts and constraints is a viable strategy for fine-tuning LLM-based question generation without retraining the underlying language model.

Table 5 reports the results of human expert evaluation of a stratified sample of generated questions. The raters judged each item on four dimensions: clarity, relevance to topic, appropriateness of difficulty, and originality. The high mean scores for clarity and topic relevance show that the LLM, when constrained with structured prompts and concept tags, can reliably produce well-formed questions that align with the intended curriculum areas. Low standard deviations indicate that the quality is consistent across items.

Table 5 **Human Evaluation of Question Quality**

Criterion	Rating Scale	Mean Score	Std. Deviation
Clarity	1–5	4.3	0.6
Relevance to Topic	1–5	4.5	0.5
Appropriate Difficulty	1–5	4.1	0.7
Originality	1–5	3.8	0.8

The slightly lower score for originality is not necessarily a negative outcome. For assessment purposes, a controlled degree of structural similarity to canonical templates is desirable, as it ensures that questions measure the same underlying construct and can be systematically calibrated. The good rating on appropriate difficulty confirms that the combination of generation workflow and difficulty scaling is able to approximate human expectations, which is critical before deploying the system in real adaptive assessment scenarios.

Difficulty Scaling and Learner Performance

Figure 7 shows a clear decreasing pattern in mean accuracy as difficulty levels escalate from Easy to Hard. This confirms that the generated difficulty tiers are behaviorally meaningful when tested with real learners. The Easy questions yield an 87% mean accuracy, reflecting high accessibility and limited conceptual complexity. This performance window is suitable for early-stage diagnosis, confidence-building, and reinforcement of basic recall skills.

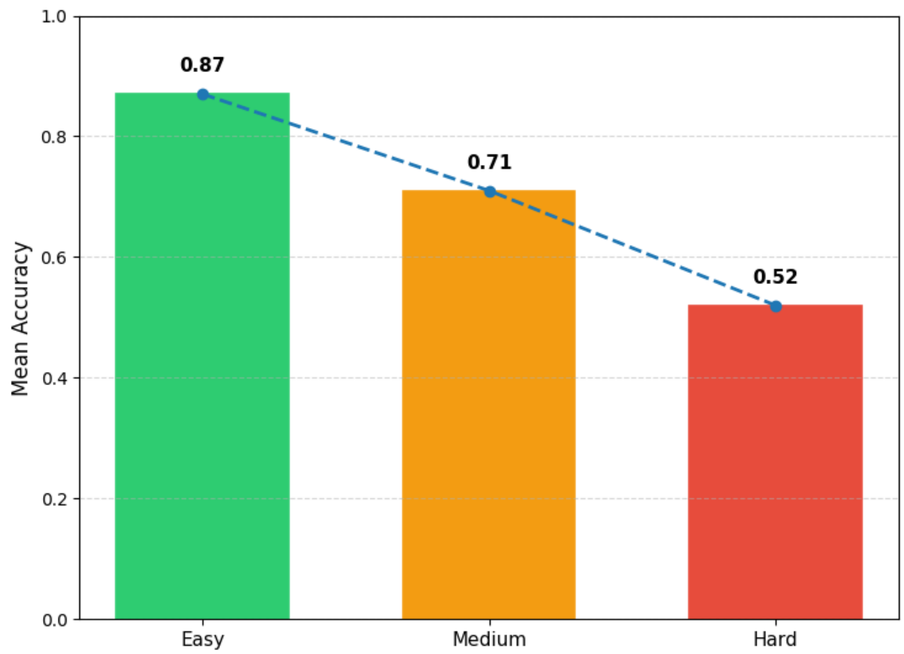

Figure 7 Learner Accuracy Across Difficulty Levels

In the Medium and Hard tiers, the average accuracy drops, indicating that the questions require more analytical reasoning and procedural fluency. A 52% mean accuracy in the Hard band is not alarming; rather, it demonstrates the system's ability to expose performance gaps that can guide subsequent recommendation. The descending trend provides behavioral validation that difficulty scaling is functioning as intended, rather than being an artificial numerical label.

Table 6 highlights how learner accuracy varies not only by difficulty band but also by domain topic. Arithmetic remains the most stable topic across difficulty levels, with accuracy dropping from 91% to 63%. The decline is measurable but not extreme, which suggests that most learners possess procedural comfort even when numerical reasoning becomes more layered. This also indicates that arithmetic-based difficulty relies more on parameter scaling than conceptual transformation.

Table 6 Performance Variation by Topic and Difficulty

Topic	Easy Accuracy	Medium Accuracy	Hard Accuracy
Arithmetic	0.91	0.78	0.63
Algebra	0.85	0.66	0.47

Geometry	0.88	0.70	0.50
Statistics	0.86	0.71	0.49

Algebra and Statistics show more aggressive performance degradation. Hard Algebra items record only 47% accuracy consistent with the topic's inherent abstraction and symbolic reasoning requirements. Statistics also falls towards the lower end because interpretation-based questions rely on contextual understanding and comparison, rather than formulaic execution. From a system perspective, this differentiation supports the adaptive engine: future item assignment can prioritize weaker domains rather than simply escalating difficulty uniformly.

Figure 8 shows performance improvement after three rounds of adaptive sequencing. In Round 1, the mean score is relatively modest at 0.64, reflecting uncertainty and early-stage calibration of ability. By Round 2, performance increases significantly, indicating that the question assignment mechanism is successfully matching difficulty to the learner's evolving competence. The jump here is aligned with the principle of "instructional fit," where correct challenge exposure leads to rapid development.

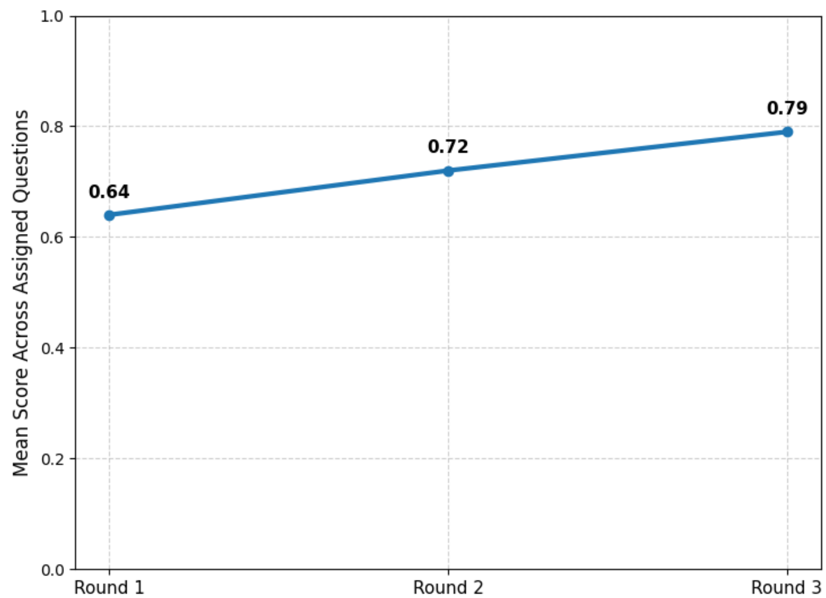

Figure 8 Learner Progression Over Three Adaptive Rounds

The continued improvement into Round 3 demonstrates that the system does not merely repeat prior content but successfully maintains pedagogical pressure. In an adaptive context, this pattern is more informative than static scores, because it highlights responsiveness to algorithmic adjustments. A flat or declining curve would indicate overshooting difficulty or insufficient variation. Instead, the upward stabilization suggests a good balance between reinforcement and stretch-target exposure.

Table 7 reports subjective learner feedback concerning difficulty alignment. The strongest response category (82% agreement) indicates that the majority perceive questions as matched to their current competence. This subjective validation aligns with the behavioral evidence shown earlier. When approximately one in five learners expresses concern about overwhelming difficulty, this is not inherently negative; it signals that the hard tier is indeed exposing cognitive friction, which is part of measurement validity.

Table 7 **Learner Feedback on Difficulty Appropriateness**

Feedback Category	Agreement Rate	Comments
Difficulty feels matched to my level	82%	Most learners reported steady challenge
Hard questions are too overwhelming	21%	A minority indicated cognitive overload
Question sequence motivates improvement	77%	Users perceived constructive discomfort

The 77% agreement rate for motivation reinforces the educational benefit of controlled difficulty escalation. Learners are not simply completing tasks they are feeling guided through a process in which success and struggle are intentionally balanced. This emotional response is relevant because adaptive testing that fails motivationally can collapse in engagement long before psychometric failure.

Conclusion

The implementation of an adaptive question generation and difficulty-scaling system powered by Large Language Models demonstrates that automated assessment can achieve both linguistic quality and pedagogical relevance when supported by structured constraints. The methodology used in this research (seed-based concept anchoring, iterative LLM refinement, and quality screening) ensures that generated items remain aligned with curriculum expectations rather than drifting into generic or hallucinated outputs. As evidenced by expert evaluation and topic-level reliability, the item bank maintains consistency in clarity, mathematical correctness, and conceptual fidelity, providing a stable foundation for scalable learning assessment.

From a learner-centric perspective, the adaptive difficulty pipeline successfully differentiates between performance levels and guides individuals across calibrated challenge bands. Behavioral indicators showed predictable reductions in accuracy between Easy, Medium, and Hard tiers, confirming that the difficulty scaling mechanism maps onto authentic performance variation rather than arbitrary labels. The reinforcement loop, supported by continuous logging and Bayesian updating, enabled measurable improvement across adaptive rounds—demonstrating that intelligent sequencing can enhance both engagement and mastery progression. Learner feedback further validated the alignment between perceived difficulty and instructional relevance.

Operationally, the system proved viable for live deployment, with latency measurements, caching triggers, and workload balancing ensuring usability under both off-peak and peak load. Reliability across content areas remained within acceptable psychometric thresholds, and discrimination distributions showed strong capacity to separate high-ability from low-ability learners. Future work should expand domain coverage, strengthen generation at the extreme difficulty boundaries, and incorporate downstream scoring models (e.g., IRT or G-theory) for full-scale educational validation. Overall, the findings affirm that LLM-driven adaptive assessment is not only computationally feasible but instructionally advantageous, offering a scalable blueprint for personalized learning systems.

Declarations

Author Contributions

Conceptualization: A.I. and S.Y.; Methodology: S.Y.; Software: A.I.; Validation: A.I. and S.Y.; Formal Analysis: A.I. and S.Y.; Investigation: A.I.; Resources: S.Y.; Data Curation: S.Y.; Writing Original Draft Preparation: A.I. and S.Y.; Writing Review and Editing: S.Y. and A.I.; Visualization: A.I.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G. Cooper, K.-S. Tang, and A. Fitzgerald, "Intersections of Mind and Machine: Navigating the Nexus of Artificial Intelligence, Science Education, and the Preparation of Pre-service Teachers," *J. Sci. Educ. Technol.*, vol. 34, no. 6, pp. 1255–1259, 2025, doi: 10.1007/s10956-025-10200-9.

[2] D. Bengs, U. Brefeld, U. Kroehne, and F. Zehner, "Joint Item Response Models for Manual and Automatic Scores on Open-Ended Test Items," *Psychometrika*, vol. 90, no. 4, pp. 1346–1367, 2025, doi: 10.1017/psy.2025.10018.

[3] S. Feng, H. Zhang, and D. Gašević, "Mapping the evolution of AI in education: Toward a co-adaptive and human-centered paradigm," *Comput. Educ. Artif. Intell.*, vol. 9, no. December, p. 100513, 2025, doi: 10.1016/j.caeai.2025.100513.

[4] Y. Zhuang, R. Zhao, Z. Xie, and P. L. H. Yu, "Enhancing language learning through generative AI feedback on picture-cued writing tasks," *Comput. Educ. Artif. Intell.*, vol. 9, no. December, p. 100450, 2025, doi: 10.1016/j.caeai.2025.100450.

[5] J. Tsao, "Trajectories of AI policy in higher education: Interpretations, discourses, and enactments of students and teachers," *Comput. Educ. Artif. Intell.*, vol. 9, no. December, p. 100496, 2025, doi: 10.1016/j.caeai.2025.100496.

[6] M.-J. Luo et al., "A large language model digital patient system enhances ophthalmology history taking skills," *npj Digit. Med.*, vol. 8, no. 1, p. 502, 2025, doi: 10.1038/s41746-025-01841-6.

[7] S. Abbasi and A. M. Rahmani, “Context-Aware Prompt Engineering for Large Language Models in Autonomous Vehicles,” *Concurr. Comput. Pract. Exp.*, vol. 37, no. 27–28, p. e70392, 2025, doi: 10.1002/cpe.70392.

[8] Y. Liu, W. Yuan, W. Chen, W. Li, H. Yang, and Y. Zhang, “CPLLM-WPF: A multi-scale prompting framework for generalizable wind power forecasting with LLMs,” *Appl. Energy*, vol. 402, no. December, p. 126912, 2025, doi: 10.1016/j.apenergy.2025.126912.

[9] N. Jeršič, M. Turkanović, and T. Beranic, “Towards a Sustainable Cybersecurity Governance: Threat Modelling with Large Language Models,” *Sustain.*, vol. 17, no. 23, p. 10569, 2025, doi: 10.3390/su172310569.

[10] M. Al-Olaqi, A. Al-Gailani, and M. M. H. Rahman, “Comprehensive Study of SQL Injection Attacks Mitigation Methods and Future Directions”, *Journal of Cyber Security and Risk Auditing*, vol. 2025, no. 4, pp. 347–365, 2025, doi: 10.63180/jcsra.thestap.2025.4.11.

[11] M. Doležel and R. Lískovec, “Reference and Solution Architecture for GenAI- and GIS-Enhanced Physical Activity Interventions: Towards Implementing the AI4Motion Platform,” *J. Med. Syst.*, vol. 49, no. 1, p. 150, 2025, doi: 10.1007/s10916-025-02269-x.

[12] M. Wang, Y. Shen, B. Zhao, X. Zhou, L. Sun, and X. Liu, “Enhancing LLM-based clinical reasoning in anesthesiology via graph-augmented retrieval and explainable generation,” *Heal. Inf. Sci. Syst.*, vol. 13, no. 1, p. 62, 2025, doi: 10.1007/s13755-025-00379-x.

[13] R. Ghazawi and E. Simpson, “How well can LLMs grade essays in Arabic?,” *Comput. Educ. Artif. Intell.*, vol. 9, no. December, p. 100449, 2025, doi: 10.1016/j.caeai.2025.100449.

[14] A. Alshehri, “Developing a multi-layer agent framework to enhance AI-generated educational questions for cybersecurity,” *J. Umm Al-Qura Univ. Eng. Archit.*, vol. 16, no. 4, pp. 1045–1056, 2025, doi: 10.1007/s43995-025-00136-x.

[15] D. Jiang, Y. Lan, and G. Yue, “Adaptive boosting estimation algorithms on bond of steel–concrete in steel reinforced composite structures,” *Signal, Image Video Process.*, vol. 19, no. 13, p. 1073, 2025, doi: 10.1007/s11760-025-04592-9.

[16] Y. Yang, P. Wang, X. Liu, W. Luo, and L. Yang, “A Multi-Agent and GraphRAG-Based Framework for Operation and Management Decision-Making in Hydraulic Projects,” *Water Resour. Manag.*, vol. 39, no. 14, pp. 7665–7687, 2025, doi: 10.1007/s11269-025-04312-5.

[17] F. Cornillie, J. Gijpen, S. Said-Metwaly, S. Luypaert, and W. van den Noortgate, “Toward Adaptive Spoken Dialogue Systems for Language Learning: Predicting Task Completion from Learning Process Data,” *CALICO J.*, vol. 42, no. 3, pp. 413–436, 2025, doi: 10.3138/calico-2025-0035.

[18] P. García, J. de Curtò, I. de Zarzà, J. C. Cano, and C. T. Calafate, “Foundation Models for Cybersecurity: A Comprehensive Multi-Modal Evaluation of TabPFN and TabICL for Tabular Intrusion Detection,” *Electron.*, vol. 14, no. 19, p. 3792, 2025, doi: 10.3390/electronics14193792.

[19] X. Chen, “Sustainable Agile Identification and Adaptive Risk Control of Major Disaster Online Rumors Based on LLMs and EKGs,” *Sustain.*, vol. 17, no. 19, p. 8920, 2025, doi: 10.3390/su17198920.

[20] Z. Wei et al., “Advanced Smart Contract Vulnerability Detection via LLM-Powered Multi-Agent Systems,” *IEEE Trans. Softw. Eng.*, vol. 51, no. 10, pp. 2830–2846, 2025, doi: 10.1109/TSE.2025.3597319.

[21] S. Hu et al., “Masterpiece Creation: An AI-powered Robotic Calligraphy Creation System,” *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 9, no. 3, pp. 1-34, 2025, doi: 10.1145/3749957.

[22] M. Zając, “Heuristic, Hybrid, and LLM-Assisted Heuristics for Container Yard Strategies Under Incomplete Information: A Simulation-Based Comparison,” *Appl. Sci.*, vol. 15, no. 18, p. 10033, 2025, doi: 10.3390/app151810033.

[23] M. Li, J. Lin, and Z. Yang, “Confidence-guided Prompt Learning for Multimodal Aspect-level Sentiment Analysis,” *Comput. Sci.*, vol. 52, no. 7, pp. 241–247, 2025, doi: 10.11896/jsjkx.240600126.

[24] M.-Y. Du et al., “Constructing Benchmark Datasets for Privacy-Protected User Comments and Evaluating the Reasoning Capabilities of Large Models,” *Jisuanji Xuebao/Chinese J. Comput.*, vol. 48, no. 7, pp. 1530–1550, 2025, doi: 10.11897/SP.J.1016.2025.01529.

[25] J. Deng, C. Yang, L. Ren, X. Li, C. Wang, and Z. Bai, “A review on coal spontaneous combustion prediction based on machine learning,” *Meitan Xuebao/Journal China Coal Soc.*, vol. 50, pp. 336–360, 2025, doi: 10.13225/j.cnki.jccs.2024.1369.

[26] D. He, H. Pu, and J. He, “Venous Thrombosis Risk Assessment Based on Retrieval-Augmented Large Language Models and Self-Validation,” *Electron.*, vol. 14, no. 11, p. 2164, 2025, doi: 10.3390/electronics14112164.