ISSN 2522-6398 Vol. 8, Issue 12: 78-85, DOI: 10.25236/FER.2025.081212

Generative AI-Based System for Automatic Exercise Generation in Middle School Mathematics: Design and Evaluation Methodology

Yan Bingpeng[1,a,*], Luo Meifen[1], Qu Yujing[1]

[1]*College of Computer and Information Science, Chongqing Normal University, Chongqing, China*
[a]*2911485269@qq.com*
[*]*Corresponding author*

Abstract: *With the rapid advancement of artificial intelligence (AI), generative AI technologies are increasingly applied in education. This study focuses on middle school mathematics, proposing a systematic methodology integrating cognitive quantitative analysis and multi-phase quality assurance. Leveraging the DeepSeek-7B model, we construct a dynamic cognitive load quantification framework to achieve dynamic alignment between question difficulty and student cognitive profiles. An evaluation system spanning three dimensions—question quality, cognitive adaptability, and instructional practicality—is established to systematically analyze the efficacy of generative AI in automatic exercise generation. The methodology effectively validates output quality and offers novel insights for educational technology by optimizing cognitive load alignment. Future research could explore multimodal input optimization and real-time compensatory mechanisms to further enhance generative performance.*

Keywords: *Generative AI; Middle School Mathematics; Cognitive Load; Automatic Exercise Generation*

1. Introduction

Artificial intelligence (AI), as a pivotal driver and strategic technology in the new wave of scientific and industrial transformation, is accelerating its deep integration with education, driving the transition toward a digitized and intelligent educational paradigm[1]. In the educational domain, particularly in automatic exercise generation, generative AI demonstrates immense potential. From a learning theory perspective, generative AI's optimization of learning processes aligns with the information-processing logic of Gagn é's Nine Events of Instruction. According to this theory, AI-generated content can effectively support students' cognitive construction through a three-phase approach: presenting stimuli, providing guidance, and facilitating transfer[2]. Concurrently, information processing theory indicates that personalized AI-generated exercises reduce extraneous cognitive load, enabling students to allocate working memory resources to core knowledge acquisition[3]. These theories provide a foundational rationale for the educational application of generative AI. Traditional exercise generation methods rely on manual teacher compilation, often resulting in outdated content, inconsistent difficulty levels, and insufficient personalization[4]. In contrast, generative AI can dynamically generate novel exercises based on input knowledge points and difficulty parameters, offering an efficient and flexible solution for educational practice. Comparison between traditional writing and AI-generated content (Table 1)

Table 1: Brief Comparison Between Traditional and AI-Based Exercise Generation Methods

Limitations of Traditional Methods Content Stagnation	Advantages of AI-Based Solutions
Manual Authoring Time by Teachers: >2 hours per set	Real-time Generation: <5 minutes
Exercise Repetition Rate: ~40%	Repetition Rate After LoRA Fine-tuning: ≤15%
Static Difficulty Adaptation	Dynamic Cognitive Load Regulation

This study innovatively employs the DeepSeek-7B model (DeepSeek-R1) as the generative engine, integrating a cognitive load assessment framework to construct an automatic generation framework that balances technical feasibility and educational appropriateness. Concurrently, we propose a

ISSN 2522-6398 Vol. 8, Issue 12: 78-85, DOI: 10.25236/FER.2025.081212

comprehensive evaluation methodology to analyze whether the system-generated exercises meet the acceptance level required by frontline educators.

2. Literature Review

2.1 Evolution of Generative AI Technologies

Generative AI can quickly learn from vast amounts of data to acquire human societal knowledge. With its powerful reasoning capabilities, it extracts knowledge and intent from human instructions, ultimately generating content and knowledge[5]. Unlike traditional AI, which focuses on recognizing and predicting data patterns, generative AI centers on learning data distributions to create new content. By leveraging deep neural architectures to comprehend latent features of input data, it synthesizes new images, text, audio, or video with emergent properties[6]. Pioneering generative AI models include the GPT series, which have achieved breakthroughs in natural language processing and multimodal generation. Among these, DeepSeek-7B, released in 2024 as an efficient open-source model, delivers 128K long-context support with merely 7 billion parameters, exhibiting superior mathematical reasoning and code generation capabilities relative to models of comparable scale. Its optimized inference efficiency makes it a critical solution for resource-constrained deployment scenarios, driving the trend toward specialized and cost-effective generative AI—making it particularly suitable for automatic exercise generation in mathematics.

2.2 Applications of Generative AI in Education

When confronted with waves of technological revolution, humanity consistently responds by leveraging technology to optimize education and, in turn, using education to advance societal progress[7]. The personalized content generation capability of generative AI enables effective cognitive load management. According to Cognitive Load Theory, AI dynamically adjusts question difficulty and presentation formats to reduce extraneous cognitive load, thereby facilitating schema construction[8]. This is particularly critical in mathematical problem-solving, where optimal allocation of cognitive resources directly impacts students' problem-solving efficiency. As a specialized application of generative AI in mathematics education, mathematical generative AI systems leverage advanced algorithms and extensive educational resource repositories to deliver highly personalized and efficient learning experiences[9]. In recent years, the educational applications of generative AI have expanded significantly: on one hand, it assists teachers in rapidly developing tailored instructional materials and exercises; on the other, it provides students with adaptive learning resources and real-time feedback. However, due to inherent limitations of AI models, several studies have identified potential logical errors or biases in generated content[10]. When processing mathematical problems containing distracting conditions, model accuracy may significantly decline. This suggests that generative AI struggles to identify and filter non-essential information—a critical challenge in geometry problems, which frequently incorporate redundant graphical elements or implicit theorem application conditions. Models tend to fall into local feature matching traps, lacking the capacity for holistic reasoning. The specific error patterns are shown in Table 2.

Table 2: Preliminary Classification of Geometric Exercise Error Patterns

Error Pattern Classification	Cause Analysis
Insufficient Conditions	e.g., omitting the core condition "diagonals bisect each other"
Logical Deduction Inconsistency	e.g., circular proof or concept equivocation
Misuse of Auxiliary Lines	e.g., adding ineffective auxiliary lines leading to increased complexity
Contextual Symbol Ambiguity	e.g., misuse of "⊥" across proof contexts

2.3 Ethical Considerations in Education

The deployment of generative AI in middle school mathematics exercise generation introduces both innovation and challenges. Algorithmic limitations may result in logical inconsistencies or knowledge misalignment, potentially undermining learning efficacy. Regional biases embedded in training data could compromise objectivity and universality, affecting educational equity. When processing extensive student behavioral data, privacy preservation requires rigorous safeguards. Overreliance on AI-generated exercises risks fostering cognitive passivity, impeding independent thinking development[11].

ISSN 2522-6398 Vol. 8, Issue 12: 78-85, DOI: 10.25236/FER.2025.081212

Furthermore, excessive educator dependence on AI tools may erode their pedagogical agency. Consequently, a balanced approach must weigh technological benefits against potential risks to ensure ethically sound implementation.

3. Research Methodology and Framework Design

3.1 Data Collection and Preprocessing

(1) We collected junior high school mathematics textbooks from a specific province, examination questions from five middle schools, and user exercise records from online platforms, with particular emphasis on regional data balance. The collected data serves as the foundation for subsequent hybrid generation strategies in automatic exercise generation.

(2) After data collection, data cleaning becomes crucial. This process exhibits multi-layered and systematic characteristics, encompassing standardized integration and quality enhancement of multi-source heterogeneous data[12].

First, a composite parsing framework is constructed by integrating Optical Character Recognition, mathematical symbol parsing, and image processing technologies. This framework performs structured transformation of raw data from textbooks, examination papers, and user logs, with parameterized processing of mathematical formulas and geometric figures to ensure machine-readability of pedagogical elements. Second, semantic analysis combined with dynamic threshold algorithms is employed for exercise deduplication and validation, eliminating redundant, invalid, or out-of-scope content[13].Through systematic data cleansing, we enhance data quality to establish a reliable foundation for subsequent DeepSeek-7B-based automatic exercise generation in middle school mathematics.

3.2 Quality Control Strategies for Exercise Generation

3.2.1 Quantification of Cognitive Load

Based on Sweller's three-dimensional cognitive load theory framework, this study introduces a dynamic weight allocation mechanism. The theory categorizes cognitive load into three types[14]:

Intrinsic Cognitive Load : Determined by the inherent complexity of learning materials and learners' prior knowledge levels. Specific indicators include the number of variables involved in problems and problem types. For example, geometric proof problems require integrating graphical elements and theorem applications, exhibiting significantly higher inherent complexity than algebraic problems.

Extraneous Cognitive Load : Unnecessary load caused by information presentation methods. This study focuses on the interference effects of redundant problem statement expressions on middle school students.

Germane Cognitive Load : Effective load promoting schema construction, including solution step guidance and prior knowledge activation intensity.

This study divides the cognitive load model into two components: base loads and modulating factors, with weight sums of 1 and additional adjustment terms respectively. The dynamic weight allocation mechanism quantifies the Cognitive Load Index (CLI) through the following formula:

Cognitive Load Index (CLI) = (50% × Step Count + 30% × Variable Quantity + 20% × Problem Description Length) × (1 – 0.15 × Prior Knowledge Level + 0.1 × Exercise Typology)

Variable Definition and Standardization:

Step Count (S, weighted 50%): Number of procedural steps required for problem-solving. Step decomposition facilitates incremental schema construction by reducing per-step cognitive load. Range: 1–10 steps; standardized as $S' = S/10$.

Variable Quantity (V, weight 30%): Number of unknowns involved in the problem. Interactions between variables significantly increase working memory load, particularly for students with low prior knowledge. Value range: 1–5 variables. Standardized as $V' = V/5$ (validated through cognitive load experiments with middle school learners).

Problem Description Length (L, weight 20%): Character count of the problem statement after tokenization. Redundant information induces attentional dispersion. Length graded on a 1–5 scale,

ISSN 2522-6398 Vol. 8, Issue 12: 78-85, DOI: 10.25236/FER.2025.081212

standardized as $L' = L/5$.

Prior Knowledge Level (K, weight %): Learner proficiency quantified via Item Response Theory (IRT) modeling, ranging 0–1 (0=no foundation, 1=proficient). Higher K-values enable procedural automation to reduce intrinsic load.

Exercise Typology (T, weight %): Binary classification variable (Algebraic problems: T=0; Geometric problems: T=1).

Formula Representation: $CLI = (0.5S' + 0.3V' + 0.2L') \times (1 - 0.15K + 0.1T)$

When K=0 and T=0 (novice learners solving algebraic problems):CLI directly reflects the intrinsic cognitive load of the exercise.

When K=1 and T=1 (proficient learners solving geometric problems):Load reduction: 15% (due to high prior knowledge)

Load increase: 10% (due to geometric complexity)Net effect: 5% reduction in overall cognitive load (CLI $\times 0.95$)

3.2.2 Geometric Rule Validation Mechanism

In geometry exercise generation, this study constructs a multi-phase verification mechanism addressing error pattern classification:1. Conditional Completeness Verification: Built on the axiomatic system of Euclidean geometry, a rule repository for core middle school theorems is established. Explicit and implicit conditions are parsed semantically and matched against theorem prerequisites.2. Logical Coherence Validation: Adirected graph model constructs condition-conclusion dependencies. Depth-first search verifies path connectivity from initial conditions to target conclusions.3. Auxiliary Line Efficacy Assessment: Based on the logical consistency principle, added lines altering graphical logic structures or introducing irrelevant geometric relations are invalidated[15].4. Dynamic Correction Mechanism: For missing explicit conditions, matched theorems are retrieved to inject necessary premises. Implicit logical discontinuities trigger question restructuring.

3.3 AI Model Generation Workflow Framework Design

A hybrid generation strategy leveraging DeepSeek, LoRA fine-tuning, and SymPy automates exercise generation through four sequential steps, figure 1 shows the prompt template for DeepSeek.

```
prompt_template

As a math teacher, generate a {question type} problem regarding 【knowledge point】 with the followin
g requirements:
- Difficulty: {difficulty level} (Level 1-5)
- Incorporate a real-life scenario, such as: {example scenario}
- Output format:
  Problem statement: [Problem text + LaTeX formulas]
  Step-by-step solution: [Step-by-step derivation, including at least 2 intermediate steps]
  Answer: \boxed{\{result\}}"""
```

Figure 1: DeepSeek Prompt Engineering Template.

(1) LoRA Fine-tuning and Rule Constraints for Reducing Repetition Rates: We implement lightweight fine-tuning on the DeepSeek-7B model using LoRA (Low-Rank Adaptation) technology. By freezing the model's core parameters and training low-rank matrices for parameter adaptation, we avoid the high computational costs associated with full-model fine-tuning[16]. Simultaneously, we integrate rule-based constraints to further optimize output quality and diversity.

(2) Constrained Decoding: Employ regular expression-based output formatting to enforce structural constraints, such as mandating the inclusion of \ boxed{} answer markers. This prevents regenerative loops caused by format-related validation failures.

(3) Post-processing: Integrates the SymPy symbolic computaton library for mathematical validation of generated answers. Incorrect exercises are automatically routed to a regeneration queue.

ISSN 2522-6398 Vol. 8, Issue 12: 78-85, DOI: 10.25236/FER.2025.081212

4. Experimental Results and Analysis Design

4.1 Generation Efficacy Analysis and Evaluation Framework

This study constructs a three-dimensional evaluation framework encompassing exercise quality, cognitive alignment, and instructional practicality through data analysis techniques including natural language processing, knowledge graph technology, and manual validation. By integrating quantitative metrics with auxiliary verification, it systematically analyzes the efficacy of generative AI in automatic mathematics exercise generation for middle schools. Evaluation results preliminarily indicate the DeepSeek-7B model's performance on key metrics, yet technical bottlenecks persist due to mathematical domain characteristics and model capability constraints.

(1) Multi-Dimensional Analysis of Generation Quality

Quality assessment of AI-generated exercises was conducted through questionnaires or interviews with multiple frontline teachers and mathematics majors. Based on their acceptance levels, the strengths and weaknesses were summarized, and the advantages of automatically generated exercises were analyzed in comparison with traditional question banks. The superiority of automatically generated exercises stems from the model's dynamic context window technology, which enables innovative question types through multimodal combinations of knowledge points[17].A knowledge graph constructed according to regional curricula and natural language processing techniques were employed to analyze the knowledge point coverage and textual grammar of DeepSeek-generated exercises[18]. The analysis confirmed that the questions essentially covered target knowledge points and their associated theorems, with correct grammar and clear semantics in applied problems.

Questionnaires were designed to collect error rate statistics for geometric proof problems and algebraic problems. Geometric problems typically involve global reasoning dependencies, which may result in higher error rates.

(2) Dynamic Adaptation of Cognitive Load

Quantitative evaluation based on Sweller's cognitive load theory demonstrates that AI-generated exercises exhibit superior alignment with middle school students' cognitive levels compared to manually authored exercises. Specifically: Through manual evaluation by multiple mathematics majors and frontline teachers, it was assessed whether DeepSeek-generated exercises effectively reduced extraneous cognitive load during problem-solving. However, the model's capability for dynamic adjustment of cognitive load may be limited. When exercises involve cross-module knowledge points, although logical accuracy is maintained, the absence of automatic difficulty-tier prompts could impose additional burden on struggling students.

(3) Feasibility Verification of Teaching Implementation

Through regional curriculum adaptability testing, this study compares the alignment of exercises generated by hybrid generation strategies, pure model generation, and manual curation with the junior high school entrance examination syllabus, analyzing the feasibility of AI-generated exercises for classroom implementation. Based on a manual review mechanism with a 1-in-20 sampling density, DeepSeek-generated exercises were audited to calculate error rates and determine whether they meet high-standard question quality criteria. While generative AI demonstrates significant cost advantages for daily student practice, special attention must be paid to comparing generation efficacy between geometry and algebra problems, as geometric exercises exhibit multiple error patterns and typically require domain expert validation[19].

4.2 Cognitive Load Optimization Verification Pathway

By adjusting scenario complexity in prompts, the proportion of high cognitive load exercises decreased. This outcome aligns with theoretical expectations from Cognitive Load Theory (CLT): Table 3 and Table 4 show the cases of algebraic problems and geometric problems respectively.

Table 3: Examples of Algebra Problems

Exercise Characteristics	re-optimization (K=0)	Post-optimization (K=1)
Problem Description	A construction team plans to complete a project in 30	Task Breakdown: ① Calculate original work efficiency ② Determine workload

ISSN 2522-6398 Vol. 8, Issue 12: 78-85, DOI: 10.25236/FER.2025.081212

	days. After working for 5 days, their efficiency increases by 20%. Calculate the actual completion time.	for the first 5 days ③ Establish equation after efficiency improvement
Number of Steps	5	3
Variable Count	3 (Total Workload, Original Efficiency, Actual Completion Days)	1 (Total Workload)
Question Stem Length	Level 4	Level 1
CLI Calculation	$(0.25+9/50+4/25)\times(1-0+0)=0.49$	$(3/20+3/50+1/25)\times(1-0.15+0)=0.25\times0.85=17/80$

Table 4: Examples of Geometric Prove Problems (T=1)

Exercise Characteristics	re-optimization (K=0)	Post-optimization (K=1)
Problem Description	In quadrilateral ABCD, AB = CD and ∠A = ∠C, prove that AD = BC.	AB = CD and ∠A = ∠C (label corresponding elements).Tasks:① Recall the SSS/SAS congruence criteria② Construct auxiliary line AE ⊥ BC③ Prove using triangle congruence
Number of Steps	6	4
Variable Count	5 (Points, Angles, Sides)	2 (Auxiliary Line, Key Side)
Question Stem Length	Level 3	Level 1
CLI Calculation	$(0.3+0.3+3/25)\times(1-0+0.1)=0.72\times1.1=0.792$	$(0.2+0.12+0.04)\times(1-0.15+0.1)=2.8\times1.05=0.378$

4.3 Exercise Innovation Expansion and Ethical Risk Considerations

Generative AI has significantly improved exercise generation efficiency in junior high school mathematics, providing strong support for student learning. However, manual inspections reveal that AI typically relies on existing templates or rules, struggling to design completely novel problem-solving approaches or question types. AI creativity is constrained by the diversity and quality of training data[20]—if certain question types or knowledge points are absent from the dataset, AI-generated content will also be limited. While enjoying its convenience and efficiency, we must also address ethical considerations in educational applications. Educational ethics constitute both a focal point and challenge when generating exercises automatically, with key priorities including ensuring fairness, accuracy, privacy protection, and teacher-technology collaboration. Measures should be taken to avoid content misinformation and cultural biases, safeguard student data privacy, and prevent excessive teacher reliance, thereby preserving educational integrity and protecting all stakeholders' rights.

5. Summary and Outlook

This study focuses on the application of generative AI in automatic exercise generation for middle school mathematics. By designing a Cognitive Load Index (CLI) standard to regulate exercise difficulty, an automatic generation framework based on the DeepSeek model was constructed, achieving notable results. In terms of generation quality, a multidimensional analysis evaluated exercise novelty and logical accuracy, with LoRA fine-tuning and rule constraints effectively reducing content repetition. The AI-generated exercises demonstrated good alignment with students' cognitive levels. While the performance of automatically generated geometric proofs fell short of expectations, the overall approach provided effective support for cognitive load reduction. Finally, through feasibility verification in teaching implementation, the study analyzed frontline teachers' acceptance rates and generation efficiency, noting its low cost and preliminary validation as a new pathway for instructional resource development.

Future research may advance in the following directions. On one hand, continuous optimization of

ISSN 2522-6398 Vol. 8, Issue 12: 78-85, DOI: 10.25236/FER.2025.081212

generative models should focus on weak areas such as geometric problems, exploring multimodal input approaches to enhance generation quality. On the other hand, the dynamic cognitive load optimization mechanism should be refined through more granular evaluation systems, enabling precise adjustment of exercise difficulty and provision of scaffolding based on students' cognitive states. Regional adaptation strategies must be developed to meet diverse curriculum requirements across regions. Additionally, ethical considerations require attention, including the design of knowledge attribution mechanisms to clarify intellectual property rights for AI-generated content, and the establishment of quality traceability systems to ensure educational safety.

This study provides valuable insights into the application of generative AI in education. Subsequent research should strengthen interdisciplinary collaboration to further unlock its educational potential, promoting deep integration of educational technology with teaching practices to contribute more significantly to the development of the education sector.

References

[1] HU H, YANG L. AI Empowering the Construction of a Country Strong in Education: Innovation in Artificial Intelligence Teaching and Application[J]. Journal of Teacher Education, 2025, 12(4): 62-70. DOI: 10.13718/j.cnki.jsjy.2025.04.007.

[2] LIN X. Gagné's Information Processing Learning Theory and Instructional Design [J]. Fujian Tribune, 2010, (S1): 100-101.

[3] YANG X., PAN Y., YAN X. Exploration of Geography Teaching Model Based on Gagné's Information Processing Learning Theory [J]. Teaching Reference of Middle School Geography, 2023, (09): 17-19+23.

[4] LI X. How to Design Primary School Math Exercises under the New Curriculum Standard [J]. Tianjin Education, 2021, (33): 112-113.

[5] LIU S., HAO X. The Challenges and Approaches of AIGC in Facilitating Educational Innovation [J]. Tsinghua Journal of Education,2024,45(03):1-12.DOI:10.14138/j.1001-4519.2024.03.000112.

[6] XU S. Generative Artificial Intelligence: Development Evolution and Industrial Opportunities [J]. AI-View, 2023, (04): 43-50. DOI: 10.16453/j.2096-5036.2023.04.005.

[7] QI Y., ZHOU H. Technology, System and Ideology: the Evolutionary Logic of Generative Artificial Intelligence in Education [J]. e-Education Research, 2024, 45(08): 28-34. DOI: 10.13811/j.cnki.eer.2024.08.004.

[8] GUO J., FENG J. Research on the Impact of Technological Burden on Primary and Secondary School Teachers under Educational Digital Transformation—Based on Cognitive Load Theory [J]. Education Review, 2025, (07): 103-112.

[9] ZHANG Q. Research on Generative Artificial Intelligence Technology Empowering Primary School Mathematics Education [J]. Intelligence, 2025, (10): 34-37.

[10] GUO G., LIANG D. Should Large Language Models Have "Intrinsic Language Generation Ability"? ——On Chomsky's Critique of ChatGPT's Limitations [J]. Studies in Philosophy of Science and Technology, 2025, 42(01): 1-9.

[11] CAI Z., DU G., YIN Y., et al. Research on AI-Assisted Student Writing Norms and Guidance Strategies Based on Educational Ethics [J]. Zhongguancun, 2025, (03): 129-131.

[12] QIAN Y., ZHANG L. Data Cleaning Technology and Its Application Based on Big Data [J]. Digital Technology & Application, 2023, 41(03): 84-86+113. DOI: 10.19695/j.cnki.cn12-1369.2023.03.25.

[13] ZHANG C., XIAO X., GU Y. Review of Data Cleaning Methods [J]. Journal of Beijing Institute of Graphic Communication, 2025, 33(03): 49-55. DOI: 10.19461/j.cnki.1004-8626.2025.03.005.

[14] ZHANG H., ZHANG D., HUANG R. The Development, Application and Reflection of Cognitive Loading Theory in the Intelligent Era——A Review for the 11th International Cognitive Load Theory Conference [J]. Modern Distance Education Research, 2018, (06): 37-44.

[15] LI L. Research on the Effective Application of Auxiliary Lines in Middle School Mathematics [J]. Mathematical and Physical Problem Solving Research, 2023, (08): 11-13.

[16] ZHU Y., GAO Y. Fine-Tuning Methods for General Large Language Models Based on LoRA and Its Variants [J]. Journal of Shanghai University of Electric Power, 2025, 41(01): 90-95.

[17] WEI Y., JIA K., ZENG R., et al. AI Innovation Development and Governance Transformation under the Breakthrough Effect of DeepSeek [J]. E-Government, 2025, (03): 2-39. DOI: 10.16582/j.cnki.dzzw.2025.03.001.

[18] HAO J., NIU H., DU J., et al. Typical Applications of Knowledge Graphs and Large Language Models Empowering Education and Teaching [J]. China Modern Educational Equipment, 2025, (11): 1-5. DOI: 10.13492/j.cnki.cmee.2025.11.003.

[19] CAO J., XIAO J., CAO Y. Automatic Geometric Problem Solution by Integrating Knowledge Points

ISSN 2522-6398 Vol. 8, Issue 12: 78-85, DOI: 10.25236/FER.2025.081212

Information [J]. Journal of Chinese Information Processing, 2023, 37(10): 86-96.
[20] CHEN Y., GU J. The Creativity and Imagination of Artificial Intelligence [J]. Journal of Foshan University (Social Science Edition), 2025, 43(01): 26-33. DOI: 10.13797/j.cnki.jfosu.1008-018x.2025.0015.