BEYOND PROBLEM-SOLVING: A MULTI-AGENT FRAMEWORK AND PEDAGOGICALLY-ALIGNED BENCHMARK FOR EVALUATING LLMS' EDUCATIONAL QUESTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have achieved impressive performance in educational applications, particularly in solving mathematical problems. However, existing evaluation benchmarks overwhelmingly emphasize solution correctness while neglecting the equally important capability of generating high-quality educational questions. This results in a fundamental gap in assessing LLMs' pedagogical competence and prevents a complete teaching–learning evaluation loop. We address this gap by introducing a multi-agent mathematics Question Generation (QG) framework that explicitly decomposes teacher-like behaviors—such as intent interpretation, curriculum planning, difficulty calibration, and solution verification—into collaborative LLM agents, and the first pedagogically aligned, multi-dimensional benchmark specifically designed to evaluate question generation quality. The benchmark assesses generated questions along five major dimensions—Accuracy, Difficulty, Novelty, Educational Value, and Human Preference Alignment—covering 19 fine-grained indicators grounded in educational psychology and assessment theory. Experimental results show that our multi-agent framework consistently outperforms strong LLM baselines across all five dimensions, with particularly large gains in higher-order pedagogical metrics such as diagnostic value, reasoning depth, and instructional clarity. Overall, this work takes an important step toward systematic and pedagogically grounded evaluation of LLM-driven educational content generation.

1 INTRODUCTION

Large Language Models (LLMs) show strong performance in educational problem solving, particularly mathematics, driven by instruction tuning and chain-of-thought reasoning Wei et al. (2022); OpenAI (2023). Benchmarks such as MMLU Hendrycks et al. (2021) and GAOKAO-Bench Zhang et al. (2023) reinforce this progress by evaluating models as *students*—their ability to answer standardized questions accurately. However, this answer-centric view captures only part of educational intelligence: authentic instruction also requires *teachers*' ability to design high-quality questions.

Question Generation (QG) is fundamental to assessment and instructional design Heilman & Smith (2010); Du et al. (2017), demanding concept targeting, difficulty calibration, and curriculum alignment. Yet recent findings (e.g., EQGBench Zhou et al. (2025)) suggest that even strong LLMs often produce ambiguous or misaligned questions. This is because pedagogically sound QG is inherently multi-stage, involving intent interpretation, curriculum reasoning, difficulty control, and solution verification.

As shown in Figure 1(a), traditional rule-based pipelines offer limited adaptability due to handcrafted templates and rules. As shown in Figure 1(b), direct LLM generation simplifies the pipeline but typically provides only coarse pedagogical control. In contrast, as shown in Figure 1(c), we propose a teacher-inspired multi-agent + RAG framework that decomposes QG into coordinated sub-tasks and supports fine-grained pedagogical control, together with evaluation along five pedagogical dimensions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

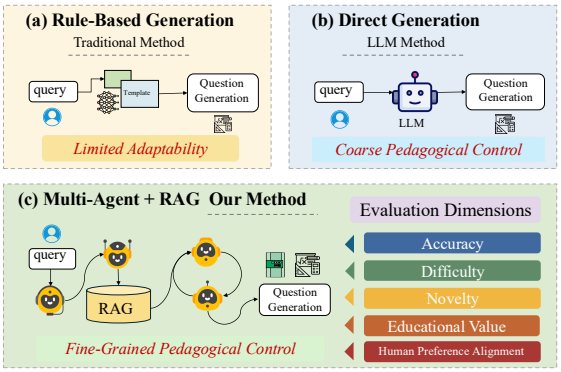


Figure 1: Comparison between previous educational math question generation paradigms and our proposed teacher-inspired multi-agent + RAG framework.

Benchmarks	V.	Size	Dom.	Ax.
LearningQ (Chen et al., 2018)	ICWSM	23K	Edu	D
QG-Bench (Ushio et al., 2022)	EMNLP	2K	Gen	E
QGEval (Fu et al., 2024)	EMNLP	3K	Gen	R
Dr.Academy (Chen et al., 2024b)	ACL	1K	Edu	T
EQGBench (Zhou et al., 2025)	arXiv	900	Edu	P
PedQG-Bench (Ours)	–	**1000**	**Edu(Math)**	**P(5)**

Table 1: Representative benchmarks for educational question generation and evaluation. **Dom.**: Edu=educational; Gen=general. **Ax.**: D=data-centric; E=extrinsic benchmarking; R=rubric-based evaluation; T=taxonomy-guided; P=pedagogy-grounded.

Despite growing interest in QG, existing benchmarks remain largely answer-centric and lack principled criteria for pedagogical quality, limiting reliable comparison across systems. To enable rigorous and pedagogy-aligned assessment, we construct a dedicated benchmark and systematically compare our multi-agent framework against strong baselines.

Our contributions are:

- We propose a teacher-inspired multi-agent framework for educational math question generation.
- We introduce a pedagogy-grounded benchmark that evaluates generated questions along five dimensions: Accuracy, Difficulty, Novelty, Educational Value, and Human Preference Alignment.
- We conduct systematic comparisons against strong baselines and provide analyses of LLMs' teacher-like QG capability.

2 Related Work

2.1 Question Generation

Early question generation (QG) relied on hand-crafted rules and syntactic transformations with statistical ranking Heilman & Smith (2010). Neural sequence-to-sequence models later enabled end-to-end QG, especially in reading comprehension Du et al. (2017); Zhao et al. (2018). More recently, instruction-following LLMs made QG flexible via prompting, and alignment methods such as RLHF improve intent following Ouyang et al. (2022). However, educational QG requires additional control over concept targeting, cognitive demand, and assessment intent that general-purpose prompting alone does not guarantee.

2.2 Educational and Math QG Benchmarks

Existing math benchmarks have substantially advanced LLM reasoning, but most emphasize *solving* rather than *generating* assessment items: chain-of-thought prompting improves solution quality Wei et al. (2022), yet does not ensure that generated problems are well-posed, curriculum-aligned, or pedagogically meaningful. Recent work benchmarks question generation itself—educational QG data Chen et al. (2018), multi-dimensional QG evaluation Ushio et al. (2022); Fu et al. (2024), taxonomy-guided questioning Chen et al. (2024b), and pedagogy-grounded criteria Zhou et al. (2025)—but these largely target general QG quality or broad educational settings. Our focus differs: curriculum-aligned mathematical item generation with explicit control over knowledge point, difficulty, and question type, a multi-agent process with retrieval, assessment, and reflection, and a 19-indicator rubric over five pedagogical dimensions (Table1).

Primary Dimension	Sub-Indicator	Discriminative Characteristics	Metric Origin
Accuracy	Logical Consistency	Determines whether a question is solvable and logically closed.	Validity theory (Messick, 1995); LLM logical consistency (Liu et al., 2025b)
	Symbol Correctness	Ensures clear notation and reduces decoding load for students.	Cognitive load theory (Sweller, 1988); readability (Graesser et al., 2004)
	Concept Definition Accuracy	Prevents concept-level hallucination even when reasoning appears correct.	Fact consistency in QA (Roberts et al., 2020); hallucination survey (Ji et al., 2023)
	Knowledge Point Alignment	Ensures content matches intended learning goals.	Constructive alignment (Biggs, 1996)
Difficulty	Concept Complexity	Higher conceptual complexity increases cognitive load.	Bloom's taxonomy (Bloom, 1956); cognitive load (Sweller, 1988)
	Reasoning Depth	Measures multi-step reasoning demand; distinguishes routine vs. conceptual reasoning.	Chain-of-thought prompting (Wei et al., 2022)
	Linguistic Complexity	Reflects comprehension difficulty independent of mathematical difficulty.	Coh-Metrix readability (Graesser et al., 2011)
	Out-of-Scope Degree	Ensures difficulty aligns with curriculum progression.	Revised Bloom taxonomy (Anderson & Krathwohl, 2001)
Novelty	Novel Perspective	Indicates ability to break template-based question patterns.	Creativity theory (Amabile, 1983; Boden, 1998)
	Expression Diversity	Distinguishes superficial rewriting from genuine re-expression.	Novelty-as-dissimilarity metrics (Liu et al., 2023)
	Cross-Concept Integration	Captures creative recombination across concepts or domains.	Combinational creativity (Boden, 1998; Scotney et al., 2019)
Educational Value	Discrimination Power	Evaluates whether a question differentiates learner ability.	CTT/IRT discrimination (Lord, 1980)
	Diagnostic Capability	Supports error analysis and personalized remediation.	Formative assessment (Black & Wiliam, 1998)
	Cognitive Activation	Evaluates higher-order thinking beyond routine practice.	Higher-order cognition (Anderson & Krathwohl, 2001)
	Feedback Potential	Indicates suitability for explanation and classroom discussion.	Effective feedback theory (Hattie & Timperley, 2007)
Human Preference Alignment	Teacher Preference Alignment	Reflects whether generated questions exhibit teacher-intuitive "good item" qualities.	Human preference alignment / RLHF (Ouyang et al., 2022)
	Style Similarity	Measures consistency with teacher phrasing and structural patterns.	Style transfer (Shen et al., 2017)
	Readability	Indicates suitability for authentic classroom use.	Educational readability (McNamara et al., 2014)
	Instructional Intent Alignment	Evaluates whether the model understands the pedagogical purpose of the question.	Instructional alignment (Anderson & Krathwohl, 2001)

Table 2: Five-dimensional pedagogical QG benchmark: dimensions, sub-indicators, discriminative characteristics, and theoretical foundations.

2.3 Multi-Agent LLM Systems for Education

Recent multi-agent LLM systems decompose complex tasks into specialized roles and iterative loops. In education, such paradigms support classroom simulation Zhang et al. (2024), simulated-student evaluation Jin et al. (2025), generative student modeling Xu et al. (2024), and tutoring pipelines Chen et al. (2024a). However, they mainly optimize task success or interaction quality rather than explicit pedagogical properties such as concept precision, difficulty control, or diagnostic value. We instead make pedagogical controllability and rubric-based evaluation central design objectives.

3 Our Pedagogical QG Benchmark

We propose a pedagogy-grounded benchmark for evaluating educational math question generation beyond surface fluency. Existing QG evaluations often emphasize answerability or linguistic quality, but under-specify instructional properties such as concept targeting, cognitive demand, creativity, diagnostic value, and alignment with teacher preferences. Our benchmark covers five dimensions—**Accuracy**, **Difficulty**, **Novelty**, **Educational Value**, and **Human Preference Alignment**—each instantiated by measurable sub-indicators grounded in learning sciences and assessment theory (Table 2).

3.1 Dimension Overview

Accuracy. Accuracy measures whether a generated item is mathematically valid, logically coherent, and conceptually precise, including logical consistency, symbol/notation correctness, concept definition accuracy, and alignment with the intended knowledge point.

Difficulty. Difficulty assesses whether the cognitive challenge matches the intended curriculum level, covering conceptual complexity, reasoning depth, linguistic complexity, and out-of-scope degree.

Novelty. Novelty captures whether questions avoid template repetition and provide new perspectives or cross-concept integration.

Educational Value. Educational Value reflects pedagogical utility, including discrimination power, diagnostic capability, cognitive activation, and feedback potential.

Human Preference Alignment. Human Preference Alignment measures whether items resemble teacher-authored questions in clarity, style, and instructional intent.

3.2 Evaluation Schema and Scoring

Table 2 specifies the full evaluation schema. For each generated question i, the score on dimension m is computed as the unweighted average of its normalized sub-indicator scores:

$$S_i^{(m)} = \frac{1}{K_m} \sum_{k=1}^{K_m} \tilde{s}_{ik}^{(m)}, \quad m \in \{A, D, N, E, H\}. \tag{1}$$

Here, $S_i^{(m)}$ denotes the score of question i on dimension m (Accuracy, Difficulty, Novelty, Educational Value, and Human Preference Alignment), $\tilde{s}_{ik}^{(m)} \in [0, 1]$ is the normalized score on the k-th sub-indicator of dimension m, and K_m is the number of sub-indicators in that dimension.

Overall score. To obtain a single scalar for ranking without introducing extra hyperparameters, we treat the five dimensions as equally important and define the overall score as:

$$S_i = \frac{1}{5} \left(A_i + D_i + N_i + E_i + H_i \right). \tag{2}$$

Model-level evaluation. Given a model that generates M questions, we report the average overall score:

$$\bar{S} = \frac{1}{M} \sum_{i=1}^{M} S_i. \tag{3}$$

We also report per-dimension averages $\bar{A}, \bar{D}, \bar{N}, \bar{E}, \bar{H}$ defined analogously (e.g., $\bar{A} = \frac{1}{M} \sum_{i=1}^{M} A_i$).

4 Methodology

4.1 Overview

We propose a collaborative multi-agent framework for automatically generating examination questions. Unlike single-pass generation, which often suffers from logical drift, our approach frames question construction as an iterative state-refinement process. As illustrated in Figure 2, specialized agents collaboratively evolve the question from a conceptual plan to a verified item.

4.2 Shared Blackboard and Graph Orchestration

To enable controllable collaboration, all agents communicate via a **Shared Blackboard** $\mathcal{B}$. Rather than passing transient messages, the blackboard maintains a persistent state tuple:

$$\mathcal{B} = \langle \mathcal{T}, \mathcal{P}, \mathcal{E}, \mathcal{I}, \mathcal{F} \rangle, \tag{4}$$

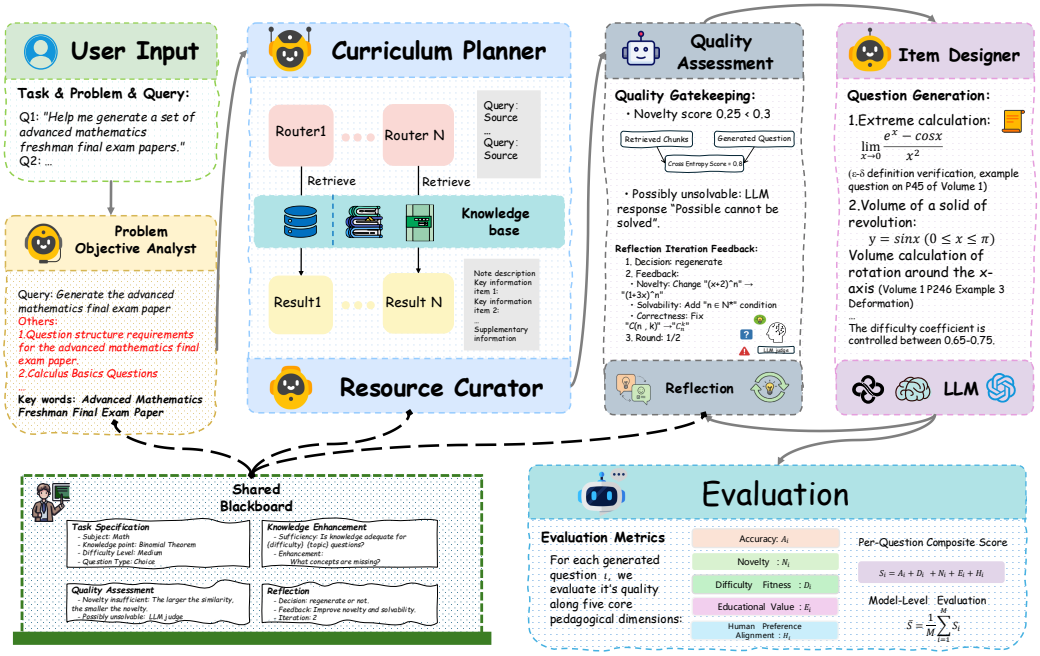


Figure 2: Overview of our LangGraph-based multi-stage multi-agent framework for automatic Mathematics question generation.

where the components denote the parsed *Task Specification* $\mathcal{T}$, the *Curriculum Plan* $\mathcal{P}$, the *Retrieved Evidence* $\mathcal{E}$ from the RAG module, the generated *Draft Item* $\mathcal{I}$, and the *Reflection Feedback* $\mathcal{F}$ produced during verification. This blackboard design offers three advantages: **state persistence** (constraints and intermediate reasoning are stored explicitly, avoiding the "catastrophic forgetting" common in long context windows); **reduced semantic drift** (agents read from and write to structured fields, keeping generation anchored to $\mathcal{P}$); and **conditional routing** (the workflow is instantiated as a LANGGRAPH whose control flow can loop back, e.g., from Refinement to Design, on the presence of $\mathcal{F}$).

4.3 AGENTS

We partition agents into three cognitive squads aligned with the problem lifecycle: *Planning*, *Synthesis*, and *Refinement*.

Problem Objective Analyst. Acting as the interface for user intent, this agent translates the raw query into a rigorous $\mathcal{T}$. It defines parameters such as the target knowledge point (e.g., Binomial Theorem), difficulty coefficient (e.g., medium), and question type (e.g., single choice) to anchor downstream generation.

Curriculum Planner. To prevent out-of-scope generation, the Planner maps the specification to a logical blueprint $\mathcal{P}$. It identifies prerequisite concepts and the expected depth of reasoning. By acting as a pedagogical scaffold, it ensures the system "thinks" before it "writes."

Resource Curator (Adaptive RAG). Responsible for evidence grounding, this agent performs retrieval using Maximal Marginal Relevance (MMR). For a query q, it selects evidence d_i to maximize diversity:

$$\text{MMR} = \arg\max_{d_i \in \mathcal{D}\setminus\mathcal{S}} \Big[\lambda \cdot \text{sim}(d_i, q) - (1-\lambda) \cdot \max_{d_j \in \mathcal{S}} \text{sim}(d_i, d_j)\Big] \tag{5}$$

The selected $\mathcal{E}$ serves as a reliable reference context to mitigate hallucinations.

Backbone Model	Acc.	Diff.	Nov.	Ed.V.	H.P.	Average
Qwen3-Max Alibaba Cloud (2025b)	4.92±0.31	3.91±0.74	1.99±0.76	3.60±0.81	4.56±0.27	3.89±0.45
+ Our Agent	**4.92±0.27**	**4.04±0.69**	**2.37±0.91**	**3.72±0.79**	**4.58±0.29**	**4.01±0.46**
Improvements	-	↑0.13	↑0.38	↑0.12	↑0.02	↑0.12
Qwen-Math-Plus Alibaba Cloud (2025a)	4.88±0.38	3.82±0.71	1.83±0.71	3.44±0.82	4.52±0.29	3.80±0.45
+ Our Agent	4.78±0.48	**4.00±0.68**	**2.16±0.77**	**3.51±0.83**	4.45±0.44	**3.86±0.48**
Improvements	↓0.10	↑0.18	↑0.33	↑0.07	↓0.07	↑0.06
Qwen3-30B-A3B Yang et al. (2025)	4.83±0.47	3.91±0.76	2.03±0.83	3.50±0.87	4.48±0.41	3.84±0.50
+ Our Agent	**4.84±0.44**	**4.08±0.67**	**2.39±0.90**	**3.64±0.85**	**4.50±0.39**	**3.97±0.51**
Improvements	↑0.01	↑0.17	↑0.36	↑0.14	↑0.02	↑0.13
DeepSeek-v3.2 Liu et al. (2025a)	4.86±0.40	**4.09±0.73**	2.32±0.89	**3.74±0.81**	**4.56±0.33**	4.00±0.48
+ Our Agent	**4.91±0.30**	4.06±0.68	**2.36±0.93**	3.73±0.78	4.55±0.35	**4.00±0.47**
Improvements	↑0.05	↓0.03	↑0.04	↓0.01	↓0.01	-
DeepSeek-R1 Guo et al. (2025)	**4.93±0.26**	3.85±0.69	1.95±0.77	3.65±0.79	**4.55±0.22**	3.88±0.41
+ Our Agent	4.81±0.44	**4.04±0.68**	**2.38±0.98**	**3.67±0.79**	4.48±0.42	**3.95±0.47**
Improvements	↓0.12	↑0.19	↑0.43	↑0.02	↓0.07	↑0.07
GPT-5.2 OpenAI (2025)	4.97±0.17	4.08±0.71	2.23±0.87	3.76±0.79	4.62±0.22	4.02±0.45
+ Our Agent	**4.98±0.09**	**4.25±0.64**	**2.99±0.96**	**3.95±0.74**	**4.63±0.22**	**4.16±0.43**
Improvements	↑0.01	↑0.17	↑0.46	↑0.15	↑0.01	↑0.14
GPT-OSS-20B Agarwal et al. (2025)	4.83±0.45	3.93±0.69	2.00±0.76	3.47±0.83	4.48±0.37	3.84±0.46
+ Our Agent	**4.88±0.41**	**4.10±0.68**	**2.38±0.90**	**3.65±0.83**	**4.53±0.35**	**3.99±0.49**
Improvements	↑0.05	↑0.17	↑0.38	↑0.18	↑0.05	↑0.15
Gemini3-Pro-Preview Google (2025)	4.98±0.13	4.09±0.66	2.28±0.88	3.79±0.72	4.62±0.19	4.04±0.42
+ Our Agent	**4.98±0.09**	**4.19±0.67**	**2.55±0.96**	**3.84±0.70**	**4.64±0.20**	**4.12±0.44**
Improvements	-	↑0.10	↑0.27	↑0.05	↑0.02	↑0.08
Gemma3-27B Gemma Team et al. (2025)	4.84±0.41	4.01±0.67	2.03±0.73	3.55±0.75	**4.52±0.29**	3.88±0.40
+ Our Agent	**4.84±0.40**	**4.03±0.68**	**2.33±0.94**	**3.66±0.79**	4.49±0.40	**3.95±0.46**
Improvements	-	↑0.02	↑0.30	↑0.11	↓0.03	↑0.07

Table 3: Performance comparison. (Acc.: Accuracy, Diff.: Difficulty, Nov.: Novelty, Ed.V.: Educational Value, H.P.: Human Preference).

Item Designer. Conditioned on the plan and evidence, this agent synthesizes the $\mathcal{I}$. We utilize template-guided generation (Stem $\rightarrow$ Conditions $\rightarrow$ Solution $\rightarrow$ Answer) to enforce structural consistency, effectively translating the abstract plan into a concrete mathematical problem.

Quality Assessor & Reflection Agent. These agents implement a verification loop. The *Assessor* evaluates the draft against solvability and clarity metrics. If the draft fails, the *Reflection Agent* diagnoses the deficiency and formulates corrective feedback $\mathcal{F}$ (e.g., "Add constraint $x > 0$"). The graph logic then routes the workflow back to the Item Designer, transforming generation into a **convergent optimization process**.

4.4 End-to-End Workflow

The overall execution follows a distinct phase-wise progression:

1. **Planning Phase:** User intent is parsed and scaffolded by the Analyst and Planner;
2. **Synthesis Phase:** The Curator retrieves evidence, and the Designer produces an initial draft;
3. **Refinement Phase:** The draft undergoes iterative critique and revision by the Assessor and Reflection Agent until quality gates are satisfied.

5 Experiments and Analysis

5.1 Experimental Setup

We evaluate our agentic framework against standalone LLMs using 9 representative backbones: Qwen3-Max Alibaba Cloud (2025b), Qwen3-30B-A3B-Instruct-2507 Yang et al. (2025), Qwen-Math-Plus Alibaba Cloud (2025a), DeepSeek-R1 Guo et al. (2025), DeepSeek-v3.2 Liu et al. (2025a), GPT-5.2 OpenAI (2025), GPT-OSS-20B Agarwal et al. (2025), Gemini3-Pro-Preview Google (2025), and Gemma-3-27B Gemma Team et al. (2025), spanning frontier proprietary and open-weight models.

	Acc.	Diff.	Nov.	Ed.V.	H.P.
SC	0.962	0.941	0.887	0.918	0.901
RC	1.000	1.000	0.500	0.500	1.000

Table 4: Human validation results. SC: score-level consistency; RC: ranking-level consistency. Higher is better.

The pipeline has two phases: **Generation**, where each backbone produces mathematical questions under *Base* (direct zero-shot) and *Agent* (our framework) settings; and **Evaluation**, where Seed2.0 Pro ByteDance Seed Team (2026a;b) scores outputs under five dimensions: *Accuracy*, *Difficulty*, *Novelty*, *Educational Value*, and *Human Preference Alignment*.

We also conduct a compact human validation study: 100 generated questions from GPT-5.2, DeepSeek-R1, and GPT-OSS-20B are scored by four experienced high-school mathematics teachers using the same rubric. As shown in Table 4, automatic scores align closely with expert judgments, with score-level consistency ranging from 0.887 to 0.962.

5.2 Main Results

Table 3 summarizes the results. Overall, our framework improves average question quality for most proprietary and open-weight models, indicating broad but not uniform model-agnostic benefits.

Overall improvements and headroom. The Agent setting improves 8 of 9 backbones in average score and ties one strong baseline, with GPT-5.2 achieving the best average score of 4.16. The gains are smaller for already strong models such as DeepSeek-v3.2 and Gemini3-Pro-Preview, where Accuracy and Human Preference Alignment are already near saturation. This suggests that the agentic workflow primarily adds value in open-ended pedagogical dimensions rather than uniformly increasing every metric.

Enhanced novelty and difficulty. The largest gains appear in *Novelty* and *Difficulty*. For example, GPT-5.2 improves by $+0.46$ in *Novelty*, suggesting that the framework reduces conservative generation and encourages more challenging, less formulaic questions while largely preserving *Educational Value*.

Closing the gap for mid-sized models. GPT-OSS-20B with our framework reaches 3.99, nearly matching DeepSeek-v3.2 at 4.00, indicating that agentic orchestration can narrow the gap between smaller and larger backbones without additional training.

5.3 Human Validation

To verify that the benchmark reflects meaningful human pedagogical judgment, we compare human scores with those produced by *Seed2.0 Pro*. We use two complementary measures: *Score-level Consistency* (SC), which captures numerical closeness between human and automatic scores, and *Ranking-level Consistency* (RC), which measures whether they rank models similarly.

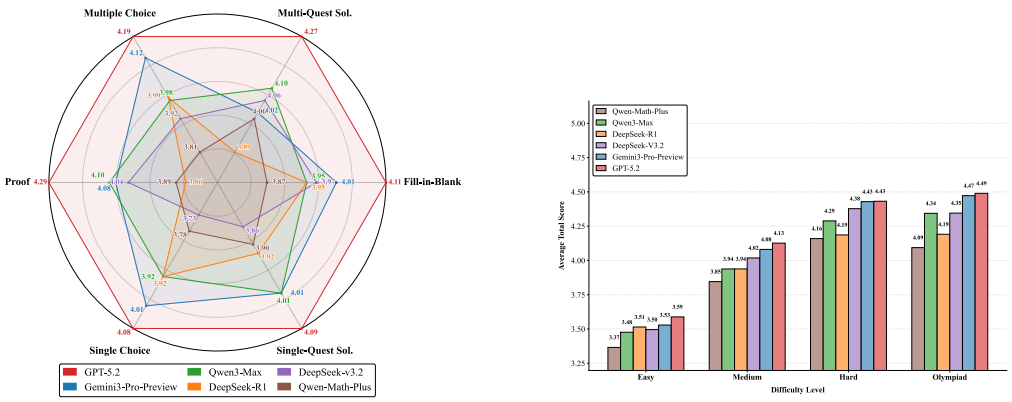
For dimension D_j, the two measures are defined as:

$$SC_j = 1 - \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N}\sum_{i=1}^{N}\left|S_k^{\text{model}}(Q_i, D_j) - S_k^{\text{human}}(Q_i, D_j)\right|\right), \tag{6}$$

$$RC_j = 1 - \frac{6\sum_{k=1}^{K} d_k^2}{K(K^2-1)}, \tag{7}$$

where $S_k^{\text{human}}(Q_i, D_j)$ denotes the average human score for question Q_i on dimension D_j for model k, $S_k^{\text{model}}(Q_i, D_j)$ is the corresponding *Seed2.0 Pro* score, $N = 100$ is the number of sampled questions, $K = 3$ is the number of compared models, and d_k is the rank difference between human and automatic evaluation for model k on dimension D_j.

Table 4 shows strong agreement between *Seed2.0 Pro* and human experts. SC ranges from 0.887 to 0.962 across all five dimensions, indicating close numerical alignment. RC is 1.000 for *Accuracy*,



(a) Question Type Performance (b) Difficulty Performance

Figure 3: Multi-granular analysis of the proposed agentic framework. (a) Aggregate performance across diverse mathematical task types. (b) Performance across four difficulty levels.

Difficulty, and *Human Preference Alignment*, and 0.500 for *Novelty* and *Educational Value*, which is expected given their greater subjectivity. Overall, these results suggest that the *Seed2.0 Pro*-based scoring pipeline is a reliable proxy for expert evaluation in large-scale benchmarking.

5.4 Fine-grained Capability Analysis

Task-Specific Performance. Figure 3a shows aggregate performance across six mathematical task types. The agentic workflow yields the largest gains on reasoning-intensive tasks, especially *Proof* and *Multi-Question Solution*, suggesting that multi-stage deliberation and verification are particularly beneficial for long-range logical synthesis.

Difficulty-wise Performance. Figure 3b shows a clear difficulty gradient that effectively differentiates model capabilities. GPT-5.2 performs best overall, reaching 4.49 in the Olympiad category, followed closely by Gemini3-Pro-Preview at 4.47. Qwen3-Max also attains 4.34 and surpasses DeepSeek-R1 (4.19) in the Olympiad setting, indicating strong reasoning ability on challenging questions. Overall, these results show that our framework generates high-quality questions across difficulty levels and serves as an effective benchmark for probing the limits of state-of-the-art models.

5.5 Ablation Study

To better understand the contribution of each core component, we conduct an ablation study using GPT-5.2 as the backbone. We compare the following settings:

- **Full System**: the complete framework with both retrieval-augmented generation (RAG) and reflection-based refinement.
- **w/o RAG**: retrieval is disabled, and the agent relies solely on the backbone's internal knowledge.
- **w/o Refinement**: the iterative reflection-and-revision loop is removed, and the system outputs the initial generation directly.

As shown in Table 5, removing either component degrades overall performance, confirming that both modules play distinct and complementary roles in the framework.

Impact of RAG. Removing RAG causes the largest drop in *Difficulty* (4.25 $\rightarrow$ 4.03), while *Accuracy* remains almost unchanged. This suggests that although the backbone can preserve mathematical correctness using internal knowledge, external pedagogical retrieval is important for calibrating the complexity of generated questions and preventing oversimplified outputs.

Impact of Refinement. The most substantial decrease under this ablation appears in *Novelty* (2.69 $\rightarrow$ 2.16). This supports our hypothesis that single-pass generation tends to exhibit instructional inertia, whereas the reflection-based refinement loop is critical for pushing the model toward more creative and pedagogically richer question formulations.

Setting	Acc.	Diff.	Nov.	Ed.V.	H.P.	Avg.
Full System	**4.98**	**4.25**	**2.69**	**3.91**	4.63	**4.16**
— w/o RAG	4.97	4.03	2.55	3.87	4.58	4.08
— w/o Refinement	4.97	4.10	2.16	3.74	**4.66**	4.02
Base Model	4.97	4.08	2.23	3.76	4.62	4.02

Table 5: Ablation results of the proposed framework. Removing key components leads to overall degradation, particularly in difficulty alignment and novelty.

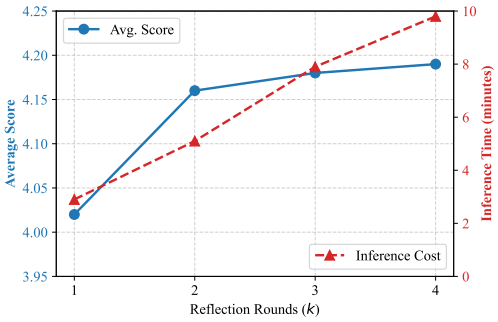


Figure 4: Performance–cost trade-off with different numbers of reflection rounds. Gains saturate beyond $k = 2$, while latency continues to increase approximately linearly.

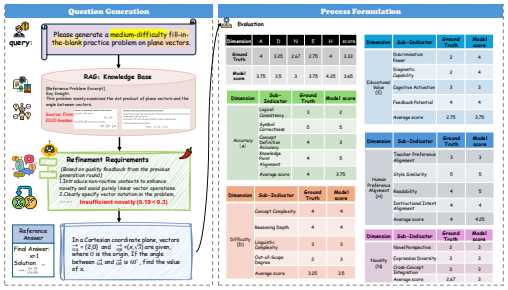


Figure 5: Case study of the agentic question generation process.

Impact of Reflection Rounds. Beyond component removal, we further investigate the sensitivity of the framework to the number of reflection rounds, denoted by k. We vary k from 1 to 4 and report both the average quality score and relative inference latency. As shown in Figure 4, increasing the number of reflection rounds initially yields a clear improvement, especially from $k = 1$ to $k = 2$. However, the benefit quickly saturates after $k = 2$, while computational cost continues to grow almost linearly. Therefore, to balance effectiveness and efficiency, we adopt $k = 2$ as the default setting in all experiments.

5.6 Case Study

Figure 5 shows a representative example of the proposed framework generating and iteratively refining a medium-difficulty plane-vector fill-in-the-blank question. The example highlights the controllability and auditability of our "generate–score–revise" loop.

6 Conclusion

We present a multi-agent framework for mathematics QG that decomposes teacher-like behaviors into collaborative agents, paired with a pedagogy-grounded rubric scoring questions along Accuracy, Difficulty, Novelty, Educational Value, and Human Preference Alignment. Experiments show broad gains, especially in novelty and difficulty, with smaller headroom for strong models near saturation, and the rubric provides interpretable signals for systematic comparison and improvement of curriculum-aligned question generation.

REFERENCES

Sandhini Agarwal et al. Gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Alibaba Cloud. Math language model - model studio help center, 2025a. URL https://www.alibabacloud.com/help/en/model-studio/math-language-model. Accessed: 2026-01-02.

Alibaba Cloud. Qwen3-max: Model studio documentation, 2025b. URL https://modelstudio.console.alibabacloud.com/?tab=/doc/?type=model&url=2840914_2&modelId=qwen3-max. Accessed: 2026-01-02.

Teresa M. Amabile. *The Social Psychology of Creativity*. Springer, 1983.

Lorin W. Anderson and David R. Krathwohl. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, 2001.

John Biggs. Enhancing teaching through constructive alignment. *Higher Education*, 32(3):347–364, 1996.

Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74, 1998.

Benjamin S. Bloom. *Taxonomy of Educational Objectives: Handbook I: Cognitive Domain*. Longmans, Green, 1956.

Margaret A. Boden. Creativity and artificial intelligence. *Artificial Intelligence*, 103(1-2):347–356, 1998. doi: 10.1016/S0004-3702(98)00055-1.

ByteDance Seed Team. Seed 2.0 official launch, 2026a. URL https://seed.bytedance.com/en/blog/seed-2-0-official-launch. Accessed: 2026-03-25.

ByteDance Seed Team. Seed 2.0, 2026b. URL https://seed.bytedance.com/en/seed2. Accessed: 2026-03-25.

Guanliang Chen et al. Learningq: A large-scale dataset for educational question generation. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pp. 481–490. AAAI Press, 2018.

Yulin Chen et al. Empowering private tutoring by chaining large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 354–364, 2024a. doi: 10.1145/3627673.3679665.

Yuyan Chen et al. Dr.academy: A benchmark for evaluating questioning capability in education for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3138–3167, 2024b. doi: 10.18653/v1/2024.acl-long.173.

Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1342–1352, 2017. doi: 10.18653/v1/P17-1123.

Weiping Fu et al. Qgeval: Benchmarking multi-dimensional evaluation for question generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11783–11803, 2024. doi: 10.18653/v1/2024.emnlp-main.658.

Gemma Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Google. Gemini 3 pro preview - vertex ai model garden, 2025. URL https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-3-pro-preview. Accessed: 2026-01-02.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202, 2004. doi: 10.3758/BF03195564.

Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234, 2011. doi: 10.3102/0013189X11413260.

Daya Guo et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77(1): 81–112, 2007.

Michael Heilman and Noah A. Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 609–617, Los Angeles, California, 2010. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Hyoungwook Jin et al. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2025. doi: 10.1145/3706598.3714054.

Aixin Liu et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025a.

Yinhong Liu et al. Aligning with logic: Measuring, evaluating and improving logical preference consistency in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 38518–38539. PMLR, 2025b.

Yiren Liu et al. Creative research question generation for human-computer interaction research. In *Joint Proceedings of the ACM IUI Workshops 2023*, volume 3359, pp. 58–66. CEUR-WS.org, 2023.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Routledge, 1980.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, 2014.

Samuel Messick. *Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning*. American Psychological Association, 1995.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774. URL https://arxiv.org/abs/2303.08774.

OpenAI. Introducing gpt-5.2, 2025. URL https://openai.com/index/introducing-gpt-5-2/. Accessed: 2026-01-02.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5418–5426, 2020. doi: 10.18653/v1/2020.emnlp-main.437.

Victoria S. Scotney, Julie McCarthy, et al. The ubiquity of cross-domain thinking in the early phase of the creative process. *Frontiers in Psychology*, 10:1426, 2019. doi: 10.3389/fpsyg.2019.01426.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6830–6841, 2017.

John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2): 257–285, 1988.

Asahi Ushio, Jose Camacho-Collados, et al. Generative language models for paragraph-level question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 670–688, 2022. doi: 10.18653/v1/2022.emnlp-main.42.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Songlin Xu et al. Eduagent: Generative student agents in learning. *arXiv preprint arXiv:2404.07963*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiaotian Zhang et al. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023. doi: 10.48550/arXiv.2305.12474. URL `https://arxiv.org/abs/2305.12474`.

Zheyuan Zhang et al. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*, 2024.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3901–3910, 2018. doi: 10.18653/v1/D18-1424.

Chengliang Zhou et al. From answers to questions: Eqgbench for evaluating llms' educational question generation. *arXiv preprint arXiv:2508.10005*, 2025. doi: 10.48550/arXiv.2508.10005.