HG-MAQG: HETEROGENEOUS GRAPH CONSTRAINED MULTI-AGENT SYSTEM FOR PERSONALIZED QUESTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality educational questions are essential for diagnosing students' knowledge states and facilitating personalized instruction, making question generation a crucial task in intelligent education systems and pedagogical agents. However, existing methods primarily emphasize linguistic fluency and template reproduction, limiting their ability to jointly support question novelty, learner-state adaptation, and cognitive progression. Recent multi-agent approaches designed to address these limitations lack a shared structured evidence base, resulting in inconsistent reasoning and error propagation across agents. To address these challenges, we propose HG-MAQG, a heterogeneous graph constrained multi-agent system for personalized question generation. HG-MAQG encodes learner states, concept prerequisites and question relations in a heterogeneous graph to ground agent reasoning in structured, shared and traceable evidence. Building on this representation, a diagnosis-driven cognitive progression mechanism uses concept-specific mastery estimates from knowledge tracing to adjust target Bloom levels, lowering cognitive demands to address mastery gaps, maintaining them to consolidate current-level competence, or raising them to introduce the next level of reasoning. The resulting cognitive targets and retrieved graph evidence jointly constrain question generation, while validation of concept coverage, cognitive alignment and novelty guides iterative refinement, with validated questions incorporated into the graph to provide additional references for subsequent generation. To evaluate cognitive alignment and learner-specific differentiation, we introduce Bloom Accuracy and Personalization Gap beyond conventional question generation metrics. Evaluation across six metrics on Eedi and XES3G5M demonstrates proposed HG-MAQG effectiveness in personalized question generation.

1 INTRODUCTION

Educational agents play a critical role in improving instructional efficiency, facilitating individualized instruction, and enhancing learning effectiveness, supporting diverse educational tasks such as adaptive learning, personalized tutoring, and learning diagnosis Zhang et al. (2026); Chu et al. (2025); Liu et al. (2024b); Gao et al. (2025); Tabarsi et al. (2025); Shi et al. (2025). Pedagogical question generation Yu et al. (2026); Lee et al. (2025); Li et al. (2025); Biancini et al. (2024); Guo et al. (2024), as a fundamental and challenging task within educational agents, has attracted increasing attention in recent years for its role in effectively guiding learners toward mastering target knowledge Liu et al. (2024a); Li et al. (2025); Hwang et al. (2024). However, designing pedagogically valuable questions requires teachers to consider student-specific knowledge states and cognitive levels Cheng et al. (2025b); Liao & Zu (2025); Li et al. (2026); Cheng et al. (2025a), imposing a substantial workload on educators. Consequently, designing pedagogical agents that generate questions adaptive to students' knowledge states while promoting their cognitive advancement remains a critical challenge that has yet to be adequately addressed.

However, limited time and effort often lead teachers to rely on one-size-fits-all massed practice and repetitive drilling, causing inefficient or even ineffective learning Shu et al. (2024); Ye et al. (2022); Ozyurt et al. (2025); Su et al. (2023). As shown in Figure 1(a), the core issue lies in mismatches between exercises and learners' differentiated cognitive needs, which manifest as insufficient current-

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.

level support, unstable same-level mastery, and limited cross-level transfer under six Bloom level and learner's Zone of Proximal Development Anderson & Krathwohl (2001); Vygotsky (1978). Thus, question generation requires jointly considering exercise personalization, learner's knowledge states and cognitive levels, rather than merely ensuring concept alignment and linguistic fluency.

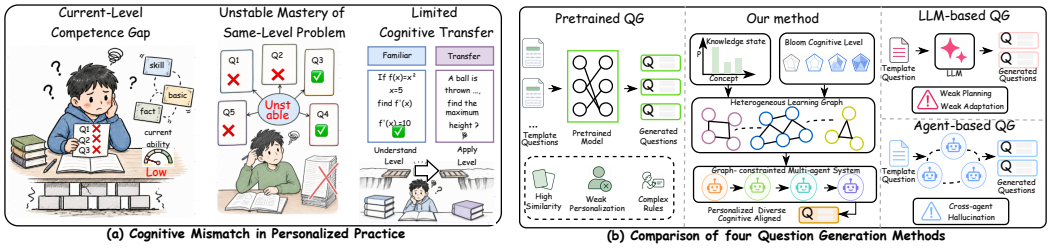


Figure 1: Illustration of cognitive mismatch and question generation challenges in personalized practice.

Existing educational question generation methods can be broadly categorized into pretraining-based, LLM-based and agent-based methods, mainly focusing on question-concept alignment and the readability and solvability of generated questions. Early methods relied on hand-crafted rules and task-specific fine-tuning with BERT-based models Ghanem et al. (2022); Rabin et al. (2023); Jia et al. (2021); Sun et al. (2022); Chen & Yen (2024); Laban et al. (2022); Yao et al. (2022); Li et al. (2022); Cui & Sachan (2023). Their dependence on rule coverage, model generalization, and annotation quality limits personalized and low-cost deployment. Advances in LLMs have demonstrated strong capabilities in reasoning and generation Yu et al. (2025); Zhao et al. (2025), while prompt-based methods reduce annotation costs and achieve promising quality Luo et al. (2024); Chen et al. (2024); Elkins et al. (2024); Li & Zhang (2024). Yet LLMs still struggle with autonomous planning and often generate template-constrained questions, limiting novelty and personalized adaptation. Recent multi-agent systems improve pedagogical alignment by decomposing question generation into collaborative stages Karbasi et al. (2025); Wu et al. (2025b); Jia et al. (2025); Wu et al. (2025b); Zhou et al. (2025); Feng et al. (2026). However, these multi-agent systems fail to translate explicit learner state modeling into cognitive targets and pedagogical strategies, and struggle to establish a shared structured evidence base for reliable collaboration among agents. To this end, as shown in Figure 1(b), current methods face three key limitations: **(1) Lack of learner-state personalization.** Students with different knowledge states require different questions for the same knowledge concept. **(2) Inadequate cognitive-level alignment and pedagogical strategy specification.** Knowledge tracing estimates are rarely translated into Bloom-level targets and generation strategies, limiting cognitively aligned remedial, consolidating, or advancing question generation. **(3) Insufficient structured representation of relations.** The intrinsic and extrinsic relations among learners, questions, and concepts are insufficiently modeled as structured evidence, leading to factual inconsistencies and limited reasoning depth in multi-agent generation Ren et al. (2024), which makes agents struggle with accurate learner diagnosis and cognitively aligned question generation.

To address these challenges, we propose a **H**eterogeneous **G**raph constrained **M**ulti-**A**gent system for personalized **Q**uestion **G**eneration (**HG-MAQG**). Specifically, (1) we develop a diagnosis-driven cognitive progression mechanism grounded in a heterogeneous graph. HG-MAQG maps mastery probabilities inferred by knowledge tracing to explicit Bloom-level targets and remediation, consolidation, or advancement strategies guided by the Zone of Proximal Development. The graph encodes relations among learners, concepts and questions as traceable evidence and determines both the target concept and the cognitive demand aligned with each learner's developmental needs. (2) We introduce a graph-constrained collaboration mechanism that coordinates agents through a shared and dynamically updated evidence base. HG-MAQG replaces unstructured textual exchanges across separate stages with an explicit cognitive progression plan and relevant graph evidence, which it propagates throughout generation and validation. The mechanism converts validation objectives, such as concept coverage, Bloom alignment, solvability and novelty, into targeted refinement signals and incorporates validated questions into the graph as new evidence. HG-MAQG establishes a closed-loop system for multi-agent question generation that integrates diagnosis, generation, validation, refinement and incremental graph updates.

In brief, our contributions can be concluded as:

(1) Diagnosis-driven cognitive progression. We formulate personalized pedagogical question generation as a diagnosis-driven task that aligns concept-specific cognitive demands with learner mastery while preserving question novelty. Using mastery estimates from knowledge tracing, HG-MAQG adjusts target Bloom levels by lowering cognitive demands to address mastery gaps, maintaining them to consolidate current-level competence, or raising them to introduce the next level of reasoning.

(2) Heterogeneous graph constrained multi-agent system. We develop a heterogeneous graph that encodes learner states, concept prerequisites, question similarities into a unified representation, providing structured, shared and traceable evidence for multi-agent reasoning. Cognitive targets and retrieved graph evidence jointly constrain generation, while validation of concept coverage, cognitive alignment and novelty guides iterative refinement, with validated questions enrich the graph as references for subsequent generation.

(3) Comprehensive experimental evaluation. We introduce Bloom Accuracy and Personalization Gap to assess cognitive alignment and learner-specific differentiation alongside Readability, Solvability, BLEU and ROUGE-L. Experiments on Eedi and XES3G5M demonstrate stronger cognitive alignment and learner-specific differentiation than existing baselines while maintaining question quality, with complementary human evaluation supporting the effectiveness of HG-MAQG.

2 PRELIMINARY

This section establishes the notation and formalizes the key concepts of personalized question generation and heterogeneous graphs.

Personalized Question Generation. Given a student's historical interaction sequence $\mathcal{H}^{\leq t} = \big((q_1, a_1), \ldots, (q_t, a_t)\big)$ and a natural-language instruction I, where $q_t \in Q$ denotes the question practiced at time t, $a_i \in \{0, 1\}$ indicates whether the student's answer is correct ($a_i = 1$) or incorrect ($a_i = 0$), and I identifies the target student for personalized question generation, the objective $J(q', c_{q'}; s_u)$ is to constraint the question q' that matches the student s_u's current knowledge $c_{q'}$ and cognitive level $l_{c_{q'}}$ while maintaining sufficient diversity from previously practiced questions to avoid repetitive practice and support cognitive progression.

Heterogeneous Graph. To support constraint-enhanced agent reasoning, we construct a heterogeneous graph $\mathcal{G}^H = (\mathcal{G}^K, \mathcal{G}^Q, \mathcal{G}^L)$, which integrates concept dependencies, question relations, and learner knowledge states into a unified, retrievable, and dynamically updatable structure. Specifically, the knowledge dependency graph $\mathcal{G}^K$ represents subject concepts and their prerequisite relations; the question graph $\mathcal{G}^Q$ models questions, their semantic representations, and their associations with relevant concepts; and the learner state graph $\mathcal{G}^L$ encodes each learner's historical interactions and current mastery over concepts, where the mastery state is estimated by a plug-and-play knowledge tracing model.

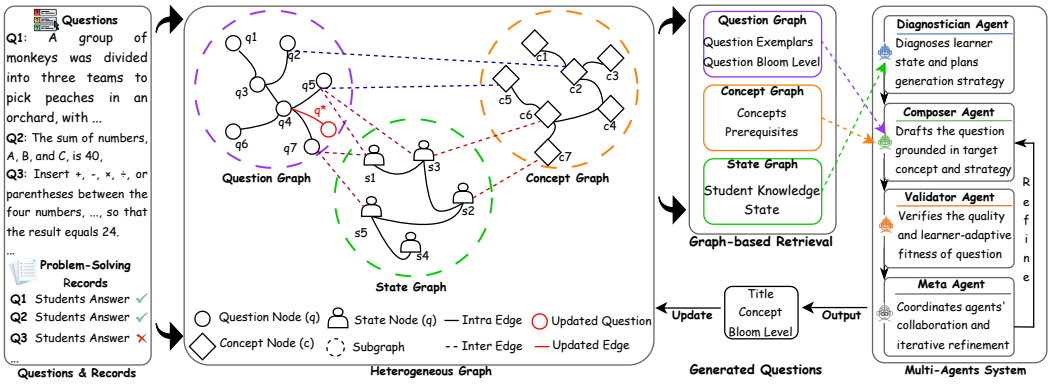


Figure 2: Overview of HG-MAQG.

3 HG-MAQG

Figure 2 illustrates how HG-MAQG adaptively designs and generates personalized questions upon the arrival of a request. Specifically, the process begins with the learner's historical learning sequence $\mathcal{H}^{\leq t}$ and the natural-language instruction I as inputs. In the graph construction stage, HG-MAQG encodes the learner's concept relations, question relations, and knowledge state into a unified heterogeneous graph $\mathcal{G}^H = (\mathcal{G}^K, \mathcal{G}^Q, \mathcal{G}^L)$. In the question generation stage, a graph-constrained multi-agent system $\mathcal{A}$, consisting of Diagnostician, Composer, Validator, and Meta Agents with clearly defined roles, conducts multi-round collaboration over the heterogeneous graph to generate personalized questions under the optimization objective constraints J .

3.1 Heterogeneous Graph Construction

As depicted in Figure 2(b), the heterogeneous graph serves as a plug-and-play module for constraint-enhanced agent reasoning, comprising a knowledge dependency graph, a question graph, and a learner state graph:

a) Knowledge Dependency Graph. The knowledge dependency graph explicitly characterizes the prerequisite relations among subject concepts covered by questions to provide knowledge-path constraints for multi-agent reasoning. The concept graph is defined as

$$\mathcal{G}^K = (C, \mathcal{E}^K)$$

where each node $c_i \in C$ represents a knowledge concept and each directed edge $(c_i, c_j) \in \mathcal{E}^K$ indicates that c_i is a prerequisite of c_j.

b) Question Graph. The graph captures semantic relations among questions and their associations with knowledge concepts, so that the multi-agent system directly retrieve relevant exemplar questions and control the diversity of generated questions. The question graph is formulated as

$$\mathcal{G}^Q = (Q, \mathcal{E}^Q, \mathcal{E}^{Q\to C}, \mathbf{X}^Q),$$

where $Q = \{q_i\}$ denotes the set of questions, and $\mathbf{X}^Q \in \mathbb{R}^{|Q|\times d}$ represents their semantic embedding matrix. Formally, the intra-question relations are derived from pairwise semantic similarities among questions by retaining the top-k nearest neighbors:

$$\mathcal{E}^Q = \{(q_i, q_j) \mid q_j \in \text{top-}k(\{sim(\mathbf{x}_i^Q, \mathbf{x}_j^Q)\}_{j\neq i})\}$$

$$sim(\mathbf{x}_i^Q, \mathbf{x}_j^Q) = \frac{(\mathbf{x}_i^Q)^T \mathbf{x}_j^Q}{\|\mathbf{x}_i^Q\|_2 \|\mathbf{x}_j^Q\|_2}$$

where $\mathcal{E}^Q$ is the set of intra-question edges, top-$k(\cdot)$ represents the k nearest neighbors, and $sim(\mathbf{x}_i^Q, \mathbf{x}_j^Q)$ denotes the semantic similarity between question embeddings $\mathbf{x}_i^Q$ and $\mathbf{x}_j^Q$.

The inter-graph edges $\mathcal{E}^{Q\to C} \subseteq Q \times C$ associate each question with its corresponding concept nodes:

$$(q_i, c_j) \in \mathcal{E}^{Q\to C} \iff c_j \in C(q_i).$$

c) Learner State Graph. The graph characterizes learners' historical response behaviors, concept mastery bloom levels, and knowledge state similarities among learners through a plug-and-play knowledge tracing model. It provides the knowledge state evidence for personalized question generation, enabling generated questions to better match learners' actual proficiency and cognitive needs. The learner state graph is formally represented as

$$\mathcal{G}^L = (S, \mathcal{E}^L, \mathcal{E}^{S\to Q}, \mathcal{E}^{S\to C}, \mathbf{M}^t),$$

where $S = \{s_u\}$ denotes the set of learners. Each learner s_u is associated with a knowledge state matrix produced by a plug-and-play knowledge tracing model:

$$m_{s_u,j,\ell}^t = f_{\text{KT}}\left(H_{s_u}^{\leq t}, c_j, \ell\right) \in [0,1], M_{s_u}^t = \left[m_{s_u,j,\ell}^t\right]_{\substack{c_j\in C \\ \ell\in\mathcal{B}}} \in (0,1]^{|C|\times|\mathcal{B}|}.$$

$$f_{\text{KT}}(H_{s_u}^{\leq t}, c_j, \ell) = \frac{\sum_{q\in Q_{s_u}^{\leq t}, c_j\in C(q)} \text{KT}(c_j, \ell, q \mid s_u)}{|\{q \in Q_{s_u}^{\leq t} \mid c_j \in C(q)\}|}$$

where $m^t_{s_u,j,\ell}$ represents the mastery degree of learner s_u over concept c_j at Bloom cognitive level ℓ at time t. $\mathrm{KT}(c_j, \ell, q \mid s_u)$ denotes the knowledge state produced by the KT model for learner s_u on concept c_j at cognitive level ℓ. The intra-learner edge set $\mathcal{E}^L \subseteq S \times S$ captures learner similarity in knowledge states:

$$\mathcal{E}^L = \{(s_u, s_v) \mid s_v \in \text{top-}k(\{sim(M^t_{s_u}, M^t_{s_r})\}_{v \neq u})\}$$

$$\mathrm{sim}(M^t_{s_u}, M^t_{s_v}) = \frac{\mathrm{vec}(M^t_{s_u})^\top \mathrm{vec}(M^t_{s_v})}{\|\mathrm{vec}(M^t_{s_u})\|_2 \|\mathrm{vec}(M^t_{s_v})\|_2}$$

where $\mathrm{vec}(\cdot)$ vectorizes the concept-by-Bloom knowledge state matrix into a single vector and the learner-question edge set $\mathcal{E}^{S\to Q} \subseteq S \times Q$ represents historical learner-question interactions:

$$(s_u, q_i) \in \mathcal{E}^{S\to Q} \iff q_i \in Q^{\leq t}_{s_u}$$

The learner-concept edge set $\mathcal{E}^{S\to C} \subseteq S \times C$ represents learners' mastery relations to concepts, with each edge associated with a Bloom specific mastery vector:

$$(s_u, c_j) \in \mathcal{E}^{S\to C}, \mathbf{w}^{S\to C,t}_{u,j} = [m^t_{s_u,j,\ell}]_{\ell \in \mathcal{B}}$$

3.2 Graph-Constrained Multi-Agent Reasoning

The Diagnostician Agent is triggered when the multi-agent system encounters a new user-provided natural-language instruction I. As the entry point of the agent system, the Diagnostician Agent first parses the instruction to identify the target learner and then retrieves the corresponding learner's knowledge state from $\mathcal{G}^L$ to return the learner's mastery evidence on the target concept together with the bloom cognitive levels of relevant historical questions:

$$D^t_{s_u,c_j} = \{(q_i,\, b(q_i),\, m^t_{s_u,j,b(q_i)}) \mid (s_u, q_i) \in \mathcal{E}^{S\to Q},\, (q_i, c_j) \in \mathcal{E}^{Q\to C}\}$$

where $D^t_{s_u,c_j}$ denotes the diagnostic evidence retrieved for learner s_u on target concept c_j at time t.

Due to sparse learners' historical response records providing partial evidence over bloom levels, directly using the observed values result in a incomplete mastery profile. To obtain a concept-specific cognitive trajectory grounded in historical evidence, we estimate missing intermediate levels by interpolating between adjacent observed levels. Specifically, for $\ell_a < \ell < \ell_b$, if $m^t_{s_u,j,\ell_a}$ and $m^t_{s_u,j,\ell_b}$ are estimated from historical responses while $m^t_{s_u,j,\ell}$ is unobserved, we define

$$m^t_{s_u,j,\ell} = \frac{\ell_b - \ell}{\ell_b - \ell_a} m^t_{s_u,j,\ell_a} + \frac{\ell - \ell_a}{\ell_b - \ell_a} m^t_{s_u,j,\ell_b}, \ell_a < \ell < \ell_b$$

This interpolation is restricted to the same concept and applied only to cognitive levels between observed levels.

Subsequently, the Composer Agent generates candidate personalized questions based on the learner diagnosis $\mathcal{D}^t_{s_u,c_j}$. Specifically, it first identifies the highest bloom level $\beta_{s_u}(c_j)$ at which the learner has available mastery evidence on concept c_j and the mastery degree $p_{s_u}(c_j)$ at level $\beta_{s_u}(c_j)$:

$$\beta_{s_u}(c_j) = \max(\mathcal{B}^{t,\mathrm{avail}}_{s_u,j}), p_{s_u}(c_j) = m^t_{s_u,j,\beta_{s_u}(c_j)}$$

Furthermore, the target cognitive level of the personalized question is determined according to three learner knowledge states. When mastery is below α, the system steps down one bloom level for remediation; when mastery falls within $[\alpha, \tau]$, it remains at the current level for consolidation; and when mastery exceeds τ, it advances one bloom level for progressive challenge. Let α be the remediation threshold and τ be the stable mastery threshold, with $0 < \alpha < \tau < 1$. The target bloom level $u_{s_u}(c_j)$ is defined as

$$u_{s_u}(c_j) = \begin{cases} \max(\beta_{s_u}(c_j) - 1, B_{\min}), & p_{s_u}(c_j) < \alpha, \\ \beta_{s_u}(c_j), & \alpha \leq p_{s_u}(c_j) \leq \tau, \\ \min(\beta_{s_u}(c_j) + 1, B_{\max}), & p_{s_u}(c_j) > \tau. \end{cases}$$

where $B_{\min}$ and $B_{\max}$ denote the lower and upper bounds of bloom levels, respectively.

Each concept is adjusted independently according to the learner's own learning state, rather than imposing a unified global bloom target across multiple concepts. Subsequently, the Composer Agent generates candidate personalized questions Q'_{s_u} conditioned on the learner diagnosis $\mathcal{D}^t_{s_u,c_j}$ and the target bloom level $u_s(c_j)$. It retrieves exemplar questions from the question graph $\mathcal{G}^Q$ by Approximate Nearest Neighbor Search (ANNS) Mishra et al. (2026) to provide question-level references that are concept-relevant and sufficiently diverse.

$$Q'_{s_u} = \text{Composer Agent}(\mathcal{D}^t_{s_u,c_j}, u_{s_u}(c_j), \mathcal{R}^t_{s_u,c_j})$$

$$R^t_{s_u,c_j} = Q^t_{s_u,c_j} \cup \{q_r \mid q_i \in Q^t_{s_u,c_j},\, q_r \in \text{ANNS}^{c_j}_{G^Q}(x^Q_i)\}$$

$$Q^t_{s_u,c_j} = \left\{q_i \;\middle|\; (s_u, q_i) \in \mathcal{E}^{S\to Q},\, (q_i, c_j) \in \mathcal{E}^{Q\to C}\right\}$$

$$\text{ANNS}^{c_j}_{G^Q}(x^Q_i) = \text{top-k}(\{\text{sim}\left(x^Q_i, x^Q_r\right) \mid q_r \in Q,\, r \neq i,\, (q_r, c_j) \in \mathcal{E}^{Q\to C}\}).$$

where $\mathcal{Q}^t_{s_u,c_j}$ denotes the historical questions practiced by learner s_u on target concept c_j, $\mathcal{R}^t_{s_u,c_j}$ denotes the exemplar question set, $\text{ANNS}^{c_j}_{\mathcal{G}^Q}(\cdot)$ returns the top-k nearest questions from the question graph $\mathcal{G}^Q$, and $\mathbf{x}^Q_i$ denotes the semantic embedding of question q_i.

The Validator Agent evaluates each candidate question through an LLM-as-a-judge procedure and deterministic post-hoc checks Zheng et al. (2023). Specifically, it derives the knowledge concepts and bloom level actually involved in the candidate question, assesses the semantic coverage of the target concepts, the alignment of bloom level and computes novelty against retrieved exemplars. These validation signals are then forwarded to the Meta Agent, which determines whether to accept the question or trigger refinement.

(O1) Bloom-level alignment. For a candidate question q' targeting concept $c_{q'}$, its Bloom level $b(q')$ matches the learner-specific target level $u_{s_u}(c_{q'})$. We define:

$$J_{\text{O1}}(q', c_{q'}; s_u) = \begin{cases} \exp(-\dfrac{(b(q') - u_{s_u}(c_{q'}))^2}{2\sigma^2}), & b(q') \le u_{s_u}(c_{q'}), \\ 0, & b(q') > u_{s_u}(c_{q'}), \end{cases}$$

where $b(q')$ denotes the bloom cognitive level of question q' and σ controls the penalty for bloom-level mismatch. This score rewards aligned questions while assigning zero reward to those exceeding the learner's current level.

(O2) Knowledge Coverage and Question Novelty. Knowledge coverage requires a candidate question to cover its target concept, while question novelty measures its distinction from exemplar questions. Following Xu et al. (2026); Zhang et al. (2024); liang et al. (2021), ROUGE-2 and ROUGE-L are used to capture both local phrase overlap and global structural similarity between candidate and exemplar questions. We define:

$$J_{\text{O2}}(q', c_{q'}; s_u) = \text{Cov}(q', c_{q'}) \cdot \text{Nov}(q'; s_u, c_{q'}).$$

$$\text{Cov}(q', c_{q'}) = \frac{|C(q') \cap C^\star|}{|C^\star|}$$

$$\text{Nov}(q'; s_u, c_{q'}) = 1 - \max_{q_e \in \mathcal{R}^t_{s_u,c_{q'}}} (\eta_2\, \text{ROUGE-2}(q', q_e) + \eta_L\, \text{ROUGE-}L(q', q_e)),$$

where $C^\star$ denotes the target concept set, $\eta_2 + \eta_L = 1$, and ROUGE-2 and ROUGE-L measure local phrase overlap and global textual structure, respectively.

The overall objective is:

$$J(q', c_{q'}; s_u) = \lambda_1 J_{O1}(q', c_{q'}; s_u) + \lambda_2 J_{O2}(q', c_{q'}; s_u)$$

The Validator Agent selects the personalized questions from the Composer Agent generated candidate set Q'_{s_u}:

$$Q^*_{s_u} = \arg \max_{\bar{Q} \subseteq Q'_{s_u},\, |\bar{Q}|=n} \sum_{q' \in \bar{Q}} J(q', c_{q'}; s_u)$$

where Q'_{s_u} denotes the candidate question set for learner s_u, $\bar{Q}$ is the selected subset, $c_{q'}$ is the target concept of q', and λ_1, λ_2 are trade-off coefficients.

Finally, the Meta Agent incrementally updates $Q^*_{s_u}$ into the question graph by connecting each accepted candidate question to its k nearest existing question nodes identified through ANNS, and sends pruned candidates with Validator feedback back to the Composer Agent for self-refinement Madaan et al. (2023) within a maximum of $iter$ rounds.

4 EXPERIMENT

In this section, we present a comprehensive evaluation of our system, assessing its performance from multiple perspectives, including: (1) The quality of personalized question generation compared with seven baseline methods across six evaluation metrics. (2) The impact of the heterogeneous graph and the graph-constrained multi-agent system on personalized question generation quality. (3) The influence of hyperparameters on the performance of HG-MAQG. (4) HG-MAQG generates personalized questions in representative cases.

4.1 EXPERIMENT SETUP

Datasets. To evaluate the effectiveness of HG-MAQG, we adopt two widely used educational knowledge tracing datasets, Eedi Wang et al. (2021) and XES3G5M Liu et al. (2023), both of which contain student response records and question texts for constructing the heterogeneous graph. We then sample 200 students from each dataset as target learners for personalized question generation. As a result, HG-MAQG generates 4,712 and 13,194 personalized questions on Eedi and XES3G5M, respectively. Details are provided in Appendix. C.1.

Baselines. We select three LLM-based baseline methods for comparison, including direct LLM prompting, Chain-of-Thought prompting Wei et al. (2022), and Dr.Academy Chen et al. (2024), denoted as Vanilla, CoT, and Academy, respectively. In addition, we adopt four agent-based methods, namely ReAct Yao et al. (2023), Socratic Holub et al. (2026), EQPR Cheng et al. (2025a), and EduAgentQG Jia et al. (2025), as representative agent-based baselines. Details are provided in Appendix. C.3.

Metrics. Readability, Solvability, BLEU and ROUGE-L are adopted as classic question generation evaluation metrics to assess the readability, solvability, and diversity of personalized questions. In addition, the proposed bloom accuracy and personalization gap are used to further evaluate the personalized suitability and personalized difference of generated questions. Details are provided in Appendix. C.2.

LLM Backbones. We adopt the open-source DeepSeek-V4-Flash model and the proprietary GPT-5.5 model as the LLM backbones for all methods, which are accessed through the DeepSeek and OpenAI APIs, respectively.

4.2 MAIN RESULTS

Table 1 reports the performance of our method and seven baselines across six evaluation metrics. The main observations are summarized as follows:

(1) HG-MAQG shows stronger personalized question generation capability. As shown in Table 1, HG-MAQG achieves the highest Bloom Accuracy scores of 0.739 and 0.741, and the highest Personalization Gap scores of 0.505 and 0.491 on Eedi and XES3G5M, respectively. Compared with the strongest baselines, it further improves Bloom Accuracy by 0.051 and 0.042, and Personalization Gap by 0.083 and 0.073. These results indicate that HG-MAQG better adapts generated questions to students with different knowledge states and Bloom-level cognitive needs.

(2) HG-MAQG generates novel questions while maintaining quality. HG-MAQG obtains the lowest BLEU scores on both datasets and the lowest ROUGE-L score on XES3G5M, indicating lower similarity to existing questions and stronger question novelty. Meanwhile, it achieves the best readability scores of 0.603 and 0.582, and strong solvability scores of 0.976 and 0.951. These results show that HG-MAQG improves question novelty while maintaining readability and solvability. More details are provided in Appendix. D.1.

Data set	Method	Read. ↑	Solv. ↑	BLEU ↓	R-L ↓	BA ↑	PG ↑
Eedi	Vanilla	0.494 ↑0.000	0.945 ↑0.000	0.509 ↑0.000	0.109 ↑0.000	0.319 ↑0.000	0.390 ↑0.000
	CoT	0.489 ↓0.005	0.935 ↓0.010	0.521 ↑0.012	0.116 ↑0.007	0.302 ↓0.017	0.347 ↓0.043
	ReAct	0.481 ↓0.013	0.953 ↑0.008	0.542 ↑0.033	0.119 ↑0.010	0.283 ↓0.036	0.385 ↓0.005
	Socratic	0.505 ↑0.011	0.946 ↑0.001	0.478 ↓0.031	0.110 ↑0.001	0.291 ↓0.028	0.422 ↑0.032
	EQPR	0.515 ↑0.021	0.925 ↓0.020	0.442 ↓0.067	0.115 ↑0.006	0.278 ↓0.041	0.318 ↓0.072
	EduAgentQG	0.498 ↑0.004	0.936 ↓0.009	0.415 ↓0.094	0.112 ↑0.003	0.290 ↓0.029	0.407 ↑0.017
	Academy	0.562 ↑0.068	0.863 ↓0.082	0.431 ↓0.078	0.130 ↑0.021	0.688 ↑0.369	0.374 ↓0.016
	HG-MAQG	**0.603** ↑0.109	**0.976** ↑0.031	**0.298** ↓0.211	0.114 ↑0.005	**0.739** ↑0.420	**0.505** ↑0.115
	HG-SAQG	0.576 ↑0.082	0.962 ↑0.017	0.386 ↓0.123	**0.105** ↓0.004	0.705 ↑0.386	0.451 ↑0.061
	w/o MA	0.577 ↑0.083	0.927 ↓0.018	0.379 ↓0.130	0.113 ↑0.004	0.662 ↑0.343	0.358 ↓0.032
	w/o HG	0.578 ↑0.084	0.944 ↓0.001	0.428 ↓0.081	0.110 ↑0.001	0.316 ↓0.003	0.347 ↓0.043
XES3G5M	Vanilla	0.480 ↑0.000	0.950 ↑0.000	0.497 ↑0.000	0.113 ↑0.000	0.303 ↑0.000	0.418 ↑0.000
	CoT	0.482 ↑0.002	0.938 ↓0.012	0.498 ↑0.001	0.112 ↓0.001	0.287 ↓0.016	0.329 ↓0.089
	ReAct	0.477 ↓0.003	0.949 ↓0.001	0.555 ↑0.058	0.116 ↑0.003	0.275 ↓0.028	0.367 ↓0.051
	Socratic	0.488 ↑0.008	0.952 ↑0.002	0.497 ↑0.000	0.113 ↓0.000	0.296 ↓0.007	0.344 ↓0.074
	EQPR	0.500 ↑0.020	0.912 ↓0.038	0.427 ↓0.070	0.108 ↓0.005	0.284 ↓0.019	0.335 ↓0.083
	EduAgentQG	0.495 ↑0.015	0.921 ↓0.029	0.429 ↓0.068	0.113 ↓0.000	0.272 ↓0.031	0.396 ↓0.022
	Academy	0.554 ↑0.074	0.850 ↓0.100	0.438 ↓0.059	0.133 ↑0.020	0.699 ↑0.396	0.363 ↓0.055
	HG-MAQG	**0.582** ↑0.102	0.951 ↑0.001	**0.306** ↓0.191	**0.098** ↓0.015	**0.741** ↑0.438	**0.491** ↑0.073
	HG-SAQG	0.573 ↑0.093	0.959 ↑0.009	0.394 ↓0.103	0.109 ↓0.004	0.711 ↑0.408	0.432 ↑0.014
	w/o MA	0.560 ↑0.080	**0.960** ↑0.010	0.415 ↓0.082	0.108 ↓0.005	0.663 ↑0.360	0.352 ↓0.066
	w/o HG	0.563 ↑0.083	0.936 ↓0.014	0.437 ↓0.060	0.107 ↓0.006	0.305 ↑0.002	0.339 ↓0.079

Table 1: Comparison of different question generation methods on Eedi and XES3G5M. The underlying LLM backbone is GPT-5.5 and we highlight the **best** and second best results. Read., Solv., R-L, BA, and PG denote Readability, Solvability, ROUGE-L, Bloom Accuracy, and Personalization Gap, respectively.

4.3 Ablation Study

As shown in Table 1, HG-SAQG performs worse than HG-MAQG, confirming the value of multi-agent collaboration. The w/o Multi-Agent variant further reduces question quality and personalized adaptation. The w/o HG variant causes the largest drops in Bloom Accuracy and Personalization Gap, showing the importance of HG for modeling students' knowledge states and Bloom-level needs. Overall, HG and Multi-Agent jointly improve question quality, personalization and novelty.

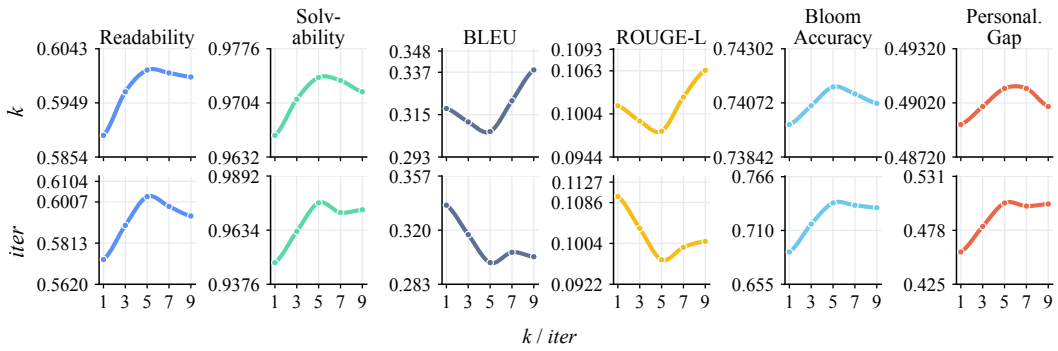


Figure 3: Impact of k and $iter$ on HG-MAQG with the Eedi datasets.

4.4 Impact of k nearest neighbors and self-refinement rounds $iter$ on HG-MAQG

As shown in Figure 3, both the number of exemplar questions k and the maximum number of self-refinement rounds iter substantially affect the performance of HG-MAQG, with the overall

optimum appearing around $k = 5$ and $\mathrm{iter} = 5$. A smaller k or iter provides insufficient retrieval guidance or self-refinement, thereby reducing generation quality and increasing BLEU and ROUGE-L scores. Further increases in k or iter lead only to slight and unstable fluctuations, indicating that both parameters require a trade-off between fully exploiting external exemplars and avoiding excessive retrieval or refinement.

4.5 Case Study

As depicted in Figure 4, HG-MAQG diagnoses Student 6 as having mastered Apply-level rectangle-area problems and advances the target to Analyze. It then generates a flower-bed expansion question requiring comparison of three area changes, while baselines mostly produce direct length, width, or area calculations. HG-MAQG achieves Bloom Match 1.0 with the highest readability, showing stronger personalization and cognitive progression.

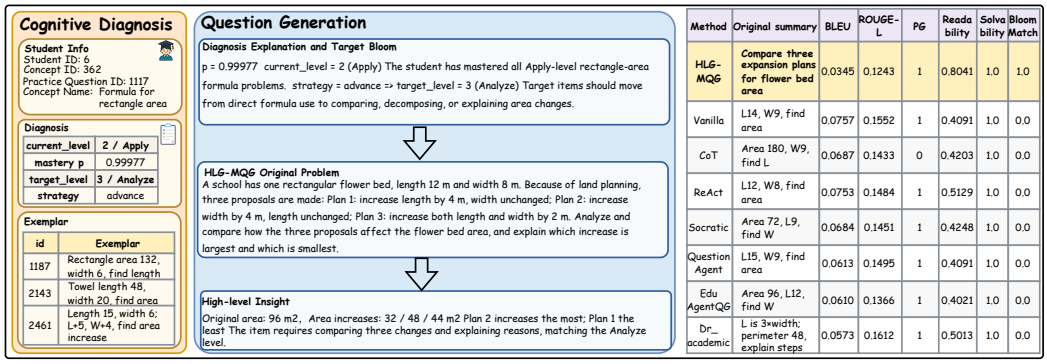


Figure 4: Case Study. L = length, W = width.

5 Related Works

Question Generation Early work treated question generation as conditional text generation using prompt engineering, knowledge injection, or fine-tuning Christ et al. (2026); Biancini et al. (2024); Chen & Yen (2024). Later studies used planning, MCTS, and knowledge-combination sampling to constrain content, objectives, and reasoning paths Cheng et al. (2025a); Jia et al. (2025); Wang et al. (2025d); Li & Zhang (2024). Other methods incorporated RAG, knowledge graphs, psychometric models, multi-agent workflows, reinforcement learning, and preference optimization to improve reliability, difficulty control, diversity, and educational effectiveness Tomikawa & Uto (2026); Wang et al. (2025c); Amanlou et al. (2026); Han Chen & Fang Shiu (2025); Jia et al. (2025); Wu et al. (2025b); Tian et al. (2026); Wei et al. (2026); Hong & Liu (2024); Ren et al. (2025).However, existing methods still lack a structured evidence-based system for personalized and novel question generation Wu et al. (2025b); Wang et al. (2025c); Wu et al. (2024); Zhou et al. (2025).

6 Conclusion

In this paper, we propose HG-MAQG, a heterogeneous graph-constrained multi-agent system for diagnosis-driven and bloom cognitively progressive personalized question generation. HG-MAQG integrates structured learner knowledge states, concept dependencies and question relations into multi-agent reason to enable adaptive cognitive level target and iterative generation under graph-based constraints. Extensive experiments and human evaluation on the Eedi and XES3G5M datasets show that HG-MAQG outperforms existing baselines across six metrics, demonstrating its effectiveness in question quality, personalization and novelty for personalized agent-driven education. For the future, the code, generated personalized questions and toolkit will be open source.

AI USE STATEMENT

In this work, we used generative AI tools as part of the multi-agent workflow for question generation. We are not used generative AI tools for any other aspects of the work. We take responsibility for the final content of this work.

ETHICS STATEMENT

The datasets used in this study were sourced from openly accessible datasets released in academic publications. During data processing, we strictly avoided the use of any personally identifiable information and the datasets do not contain inappropriate content. In addition, data annotation was conducted through a collaborative workflow using a dedicated annotation platform specifically designed for this study. The platform provided a standardized annotation interface, clear task instructions, and consistent annotation guidelines to ensure the consistency and traceability of the annotation process. All annotators were fully informed in advance about the intended use and presentation of the data, ensuring that the entire annotation process complied with relevant ethical standards.

REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of our work, we provide supplementary materials that include, but are not limited to, detailed descriptions of the datasets used in the experiments, explanations of the evaluation metrics, additional experimental analyses, the prompts used in our experiments, and the source code.

REFERENCES

Mohammad Amanlou, Erfan Shafiee Moghaddam, Yasaman Amou Jafari, Mahdi Noori, Farhan Farsi, and Behnam Bahrak. Knight: Knowledge graph-driven multiple-choice question generation with adaptive hardness calibration, 2026.

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

Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. Multiple-choice question generation using large language models: Methodology and educator insights. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '24, pp. 584–590. ACM, 2024. doi: 10.1145/3631700.3665233.

Jeanne Sternlicht Chall and Edgar Dale. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA, 1995. ISBN 9781571290083.

Mao-Siang Chen and An-Zi Yen. E-qgen: Educational lecture abstract-based question generation system. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8631–8634. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/996. Demo Track.

Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, and Yanghua Xiao. Dr.Academy: A benchmark for evaluating questioning capability in education for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3138–3167, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.173.

Cheng Cheng, Zhenya Huang, GuanHao Zhao, Yuxiang Guo, Xin Lin, Jinze Wu, Xin Li, and Shijin Wang. From objectives to questions: A planning-based framework for educational mathematical question generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12836–12856, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.628.

Weihua Cheng, Hanwen Du, Chunxiao Li, Ersheng Ni, Liangdi Tan, Tianqi Xu, and Yongxin Ni. Uncertainty-aware knowledge tracing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27905–27913, Apr. 2025b. doi: 10.1609/aaai.v39i27.35007.

Bryan R. Christ, Penelope Molitz, Beau LeBlond, Zachary Gottesman, Jonathan Kropko, and Thomas Hartvigsen. Edumath: Generating standards-aligned educational math word problems, 2026.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. LLM agents for education: Advances and applications. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13782–13810, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.743.

Peng Cui and Mrinmaya Sachan. Adaptive and personalized exercise generation for online language learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10184–10198, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.567.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Sabina Elkins, Ekaterina Kochmar, Jackie C. K. Cheung, and Iulian Serban. How teachers can use large language models and bloom's taxonomy to create educational quizzes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23084–23091, Mar. 2024. doi: 10.1609/aaai.v38i21.30353.

Yueying Feng, Wang Lin, Qingsong Wang, Yangtao Zhou, Tao Jin, Xiu Tang, Sai Wu, Chang Yao, and Jingyuan Chen. PELICAN: Personalized education via LLM-powered cognitive diagnosis and adaptive tutoring, 2026.

Rudolph Flesch. A new readability yardstick. *The Journal of Applied Psychology*, 32(3):221–233, 1948. doi: 10.1037/h0057532.

Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. Agent4Edu: Generating learner response data by generative agents for intelligent education systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23923–23932, 2025. doi: 10.1609/aaai.v39i22.34565.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, and Alona Fyshe. Question generation for reading comprehension assessment by modeling how and what to ask. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2131–2146, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.168.

Shasha Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. A survey on neural question generation: Methods, applications, and prospects. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8038–8047. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/889. Survey Track.

Ching Han Chen and Ming Fang Shiu. Kaqg: A knowledge-graph-enhanced rag for difficulty-controlled question generation. *IEEE Access*, 13:197234–197244, 2025. doi: 10.1109/ACCESS.2025.3633838.

Ondřej Holub, Essi Ryymin, and Rodrigo Alves. Reflecting in the reflection: Integrating a socratic questioning framework into automated ai-based question generation, 2026.

Zijin Hong and Jian Liu. Towards better question generation in QA-based event extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9025–9038, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.535.

Seonjeong Hwang, Yunsu Kim, and Gary Lee. Cross-lingual transfer for automatic question generation by learning interrogative structures in target languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3194–3208, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.186.

Rui Jia, Min Zhang, Fengrui Liu, Bo Jiang, Kun Kuang, and Zhongxiang Dai. Eduagentqg: A multi-agent workflow framework for personalized question generation, 2025.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. Eqg-race: Examination-type question generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13143–13151, May 2021. doi: 10.1609/aaai.v35i14.17553.

Nizam Kadir. *From Untamed Black Box to Interpretable Pedagogical Orchestration: The Ensemble of Specialized LLMs Architecture for Adaptive Tutoring*, pp. 94–109. Springer Nature Switzerland, 2026. ISBN 9783032297440. doi: 10.1007/978-3-032-29744-0_7.

Kia Karbasi, Kevin Hong, Mohammad Amin Samadi, and Gregory Pottie. Multi-agent collaborative framework for math problem generation. 2025. doi: 10.5281/ZENODO.15870246.

Jeong-Hoon Kim, Jinwoo Nam, and Geunsik Jo. Computational blueprints: Generating isomorphic mathematics problems with large language models, 2025.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. Quiz design task: Helping teachers create quizzes with automated question generation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 102–111, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.9.

Yooseop Lee, Suin Kim, and Yohan Jo. Generating plausible distractors for multiple-choice questions via student choice prediction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23669–23692, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1154.

Kunze Li and Yu Zhang. Planning first, question second: An LLM-guided method for controllable question generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4715–4729, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.280.

Yunji Li, Sujian Li, and Xing Shi. Consecutive question generation via dynamic multitask learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6620–6635, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.493.

Zhifei Li, Lifan Chen, Jiali Yi, Xiaoju Hou, Yue Zhao, Wenxin Huang, Miao Zhang, Kui Xiao, and Bing Yang. Keenkt: Knowledge mastery-state disambiguation for knowledge tracing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(23):19285–19293, Mar. 2026. doi: 10.1609/aaai.v40i23.39004.

Ziqing Li, Mutlu Cukurova, and Sahan Bulathwela. A novel approach to scalable and automatic topic-controlled question generation in education, 2025.

xiaobo liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-drop: Regularized dropout for neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10890–10905. Curran Associates, Inc., 2021.

Han Liao and Shuaishuai Zu. Routerkt: Mixture-of-experts for knowledge tracing, 2025.

Guimei Liu, Huijing Zhan, and Jung-jae Kim. Question difficulty consistent knowledge tracing. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 4239–4248, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645582.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: Exploring socratic personalized teaching with large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 85693–85721. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2721.

Yuxin Liu, Zeqing Song, Jiong Lou, Chentao Wu, and Jie Li. Agenttutor: Empowering personalized learning with multi-turn interactive teaching in intelligent education systems, 2025.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 32958–32970. Curran Associates, Inc., 2023.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7978–7993, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.432.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

Nina Mishra, Yonatan Naamad, Tal Wagner, and Lichen Zhang. Graph-based nearest neighbors with dynamic updates via random walks. In *The Fourteenth International Conference on Learning Representations*, 2026.

OpenAI. Introducing gpt-5.5, April 2026. Product release.

Yilmazcan Ozyurt, Tunaberk Almaci, Stefan Feuerriegel, and Mrinmaya Sachan. Personalized exercise recommendation with semantically-grounded knowledge tracing, 2025.

Roni Rabin, Alexandre Djerbetian, Roee Engelberg, Lidan Hackmon, Gal Elidan, Reut Tsarfaty, and Amir Globerson. Covering uncommon ground: Gap-focused question generation for answer assessment. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 215–227, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.20.

T. Ren, H. Wang, and K. Rafferty. Enhancing question generation through diversity-seeking reinforcement learning with bilevel policy decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25083–25091, 2025. doi: 10.1609/aaai.v39i23.34693.

Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh Chawla, and Chao Huang. A survey of large language models for graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6616–6626. ACM, August 2024. doi: 10.1145/3637528.3671460.

Yao Shi, Rongkeng Liang, and Yong Xu. Educationq: Evaluating llms' teaching capabilities through multi-agent dialogue framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32799–32828. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1576.

Matthew Shu, Nishant Balepur, Shi Feng, and Jordan Lee Boyd-Graber. KARL: Knowledge-aware retrieval and representations aid retention and learning in students. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14161–14178, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.784.

Jingyong Su, Junyao Ye, Liqiang Nie, Yilong Cao, and Yongyong Chen. Optimizing spaced repetition schedule by capturing the dynamics of memory. *IEEE Trans. on Knowl. and Data Eng.*, 35 (10):10085–10097, October 2023. ISSN 1041-4347. doi: 10.1109/TKDE.2023.3251721.

Jianwen Sun, Fenghua Yu, Qian Wan, Qing Li, Sannyuya Liu, and Xiaoxuan Shen. Interpretable knowledge tracing with multiscale state representation. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 3265–3276, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645373.

Yuan Sun, Sisi Liu, Zhengcuo Dan, and Xiaobing Zhao. Question generation based on grammar knowledge and fine-grained classification. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6457–6467, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

Benyamin Tabarsi, Aditya Basarkar, Xukun Liu, Dongkuan (DK) Xu, and Tiffany Barnes. MeryQuery: A trustworthy LLM-powered tool providing personalized support for educators and students. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28):29700–29702, 2025. doi: 10.1609/aaai.v39i28.35372.

Yu Tian, Linh Huynh, Katerina Christhilf, Shubham Chakraborty, Micah Watanabe, Tracy Arner, and Danielle McNamara. Cognitively diverse multiple-choice question generation: A hybrid multi-agent framework with large language models, 2026.

Yuto Tomikawa and Masaki Uto. Difficulty-controllable multiple-choice question generation using large language models and direct preference optimization. *IEEE Access*, 14:44695–44708, 2026. ISSN 2169-3536. doi: 10.1109/access.2026.3674595.

Lev Semenovich Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. Training turn-by-turn verifiers for dialogue tutoring agents: The curious case of LLMs as your coding tutors. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12416–12436, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.642.

Tianfu Wang, Yi Zhan, Jianxun Lian, Zhengyu Hu, Nicholas Jing Yuan, Qi Zhang, Xing Xie, and Hui Xiong. Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system, 2025b.

Wanqiang Wang, Longzhu He, and Wei Zheng. Beyond the textual: Generating coherent visual options for MCQs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6918–6935, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.365.

Yangfan Wang, Jie Liu, Chen Tang, Lian Yan, and Jingchi Jiang. KCS: Diversify multi-hop question generation with knowledge composition sampling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23173–23185, Suzhou, China, November 2025d. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1181.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E. Turner, Richard G. Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. Results and insights from diagnostic questions: The neurips 2020 education challenge, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022.

Yumou Wei, John Stamper, and Paulo F. Carvalho. Generate-then-validate: A novel question generation approach using small language models, 2026.

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Khoshfetrat Pakazad, Tongshuang Wu, and Graham Neubig. Synthetic multimodal question generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12960–12993, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.759.

Yang Wu, Rujing Yao, Tong Zhang, Yufei Shi, Zhuoren Jiang, Zhushan Li, and Xiaozhong Liu. Teaching according to students' aptitude: Personalized mathematics tutoring via persona-, memory-, and forgetting-aware llms, 2025a.

Zhenyu Wu, Jian Li, and Hua Huang. Magma-edu: Multi-agent generative multimodal framework for text-diagram educational question generation, 2025b.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 731–744, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.54.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Junyao Ye, Jingyong Su, and Yilong Cao. A stochastic shortest path algorithm for optimizing spaced repetition scheduling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 4381–4390, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539081.

Jifan Yu, Zheyuan Zhang, Daniel Zhang-li, Shangqing Tu, Zhanxin Hao, Rui Miao Li, Haoxuan Li, Yuanchun Wang, Hanming Li, Linlu Gong, Jie Cao, Jiayin Lin, Jinchang Zhou, Fei Qin, Haohua Wang, Jianxiao Jiang, Lijun Deng, Yisi Zhan, Chaojun Xiao, Xusheng Dai, Xuan Yan, Nianyi Lin, Nan Zhang, Ruixin Ni, Yang Dang, Lei Hou, Yu Zhang, Xu Han, Manli Li, Juanzi Li, Zhiyuan Liu, Huiqin Liu, and Maosong Sun. From mooc to maic: Reshaping online teaching and learning through llm-driven agents, 2024.

Mengxia Yu, Bang Nguyen, Olivia Zino, and Meng Jiang. Context selection and rewriting for video-based educational question generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(48):40925–40933, Mar. 2026. doi: 10.1609/aaai.v40i48.42126.

Yiyao Yu, Yuxiang Zhang, Dongdong Zhang, Xiao Liang, Hengyuan Zhang, Xingxing Zhang, Mahmoud Khademi, Hany Hassan Awadalla, Junjie Wang, Yujiu Yang, and Furu Wei. Chain-of-reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24914–24937, Vienna, Austria, July 2025.

Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1213.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Re. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *The Twelfth International Conference on Learning Representations*, 2024.

Wei Zhang, Yihang Cheng, Zhirong Ye, and Kezhen Huang. Cogevolution: A human-like generative educational agent to simulate student's cognitive evolution, 2026.

Bingxi Zhao, Jiahao Zhang, Xubin Ren, Zirui Guo, Tianzhe Chu, Yi Ma, and Chao Huang. Deeptutor: Towards agentic personalized tutoring, 2026.

Xueliang Zhao, Wei Wu, Jian Guan, and Lingpeng Kong. PromptCoT: Synthesizing olympiad-level problems for mathematical reasoning in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18167–18188, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.935.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023.

Yangtao Zhou, Hua Chu, chen chen, Ziwen Wang, Jiacheng Liu, Jianan Li, Yueying Feng, Xiangming Li, Zihan Han, and Qingshan Li. Knowledge starts with practice: Knowledge-aware exercise generative recommendation with adaptive multi-agent cooperation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 92980–93013. Curran Associates, Inc., 2025.

A Limitation and Future Work

Despite the progress made in diagnosis-driven and cognitively progressive personalized question generation, this work still has several limitations. (1) First, currently available datasets for cognitive diagnosis and personalized question generation are mainly concentrated in mathematics and are often collected from specific educational stages or teaching scenarios. This limits the model's generalization ability across different subject knowledge structures, question types, and learning behavior patterns. In future work, we plan to collaborate with real-world learning platforms and school settings to collect students' response records, knowledge-concept annotations, cognitive-level labels, and teacher feedback across multiple subjects and educational stages, thereby constructing more diverse, fine-grained, and reliable data resources. (2) Second, students' cognitive states are inherently dynamic and uncertain. Relying solely on historical response records may not fully capture short-term learning fluctuations, knowledge forgetting, strategy shifts, or the influence of non-cognitive factors such as motivation and attention. Therefore, future work will further incorporate temporal learning behavior modeling, multimodal learning-process data, and uncertainty estimation mechanisms to more accurately track changes in students' knowledge states and generate personalized exercises that are more timely, stable, and diagnostically valuable. In addition, we will open-source the code for personalized question generation, the generated personalized questions, and the related toolkit to support and inspire the community to further explore personalized question generation across a broader range of domains, rather than limiting it to educational scenarios.

B Related Work on Personalized and Adaptive Learning

In recent years, LLM-driven research on personalized and adaptive learning has gradually shifted from single-turn conversational tutoring toward systematic frameworks that integrate cognitive diagnosis, long-term memory, forgetting modeling, and multi-agent collaboration. PELICAN Feng

et al. (2026), TASA Wu et al. (2025a), ES-LLMs Kadir (2026) and TRAVER Wang et al. (2025a) enhance the tutoring process from the perspectives of collaborative cognitive diagnosis, personality and event memory, rule-based orchestration, and turn-level verification, respectively, thereby improving the awareness of students' knowledge states and learning trajectories. Meanwhile, CBIT Kim et al. (2025) and DeepTutor Zhao et al. (2026) focus on adaptive exercise generation and recommendation, supporting more fine-grained practice provision through mastery modeling, homogeneous question generation, difficulty calibration, and solution-trace grounding. Furthermore, AgentTutor Liu et al. (2025), MAIC Yu et al. (2024), and GenMentor Wang et al. (2025b) extend personalized learning to multi-turn interactive tutoring, online classrooms, and goal-oriented learning scenarios, emphasizing curriculum decomposition, learner assessment, dynamic instructional strategies, and multi-agent collaboration. These studies indicate that a key trend in LLM-based personalized education systems is to combine generative capabilities with interpretable student modeling, instructional strategy control, and adaptive resource generation. Therefore, investigating agent-based personalized question generation with LLMs as the backbone is of great significance for advancing personalized and adaptive education.

C Experimental Details

In this section, we provide further details on the datasets, evaluation metrics, and baseline methods used in our experiments.

C.1 Dataset Descriptions

In this section, we describe the datasets used in our experiments:

- **Eedi.** The Eedi dataset Wang et al. (2021) is a large-scale real-world educational dataset collected from the Eedi online learning platform. It records students' responses to multiple-choice diagnostic questions and was released for the NeurIPS 2020 Education Challenge, making it a widely used for personalized education tasks such as knowledge tracing. In our experiments, we use the Task 3 and Task 4 subsets because they provide question-stem content, including textual descriptions, rather than only anonymized question identifiers. Tasks 3 and 4 contain 948 questions and 4,918 students, with 1,382,727 response records.
- **XES3G5M.** XES3G5M is a large-scale knowledge tracing benchmark dataset Liu et al. (2023) collected from a real-world K-12 online mathematics learning platform. It contains students' response sequences as well as rich auxiliary information about questions and knowledge concepts. In our experiments, we use the full dataset to construct the heterogeneous graph and generate questions because it provides not only question identifiers but also question-stem content. The dataset contains 7,652 questions, 18,066 students, and 5,549,635 response records.

In addition, we manually annotated the Bloom cognitive levels of questions in the Eedi and XES3G5M datasets. To this end, we developed a dedicated annotation platform and invited five master's students and one doctoral student in education to complete the annotation with LLM assistance as reference. We then computed inter-annotator agreement and retained samples with Fleiss' kappa greater than 0.8. For the remaining samples with disagreement, the final labels were determined through consensus discussion. As shown in Figure 5, we will release this dedicated annotation platform to support reproducibility, extension, and community sharing in future research.

C.2 Metrics Descriptions

To evaluate personalized question generation from multiple dimensions, we use Readability to assess question readability, Solvability to assess question solvability, and BLEU and ROUGE-L to measure question diversity. We further introduce Bloom Accuracy to evaluate the alignment between generated questions and the target Bloom cognitive level, and Personalization Gap to measure the semantic differences among personalized questions generated for different students. Specifically, the evaluation metrics are computed as follows:

Readability. To comprehensively evaluate the readability of generated questions, we compute a composite readability score following the evaluation protocol of EDUMATH Christ et al. (2026),

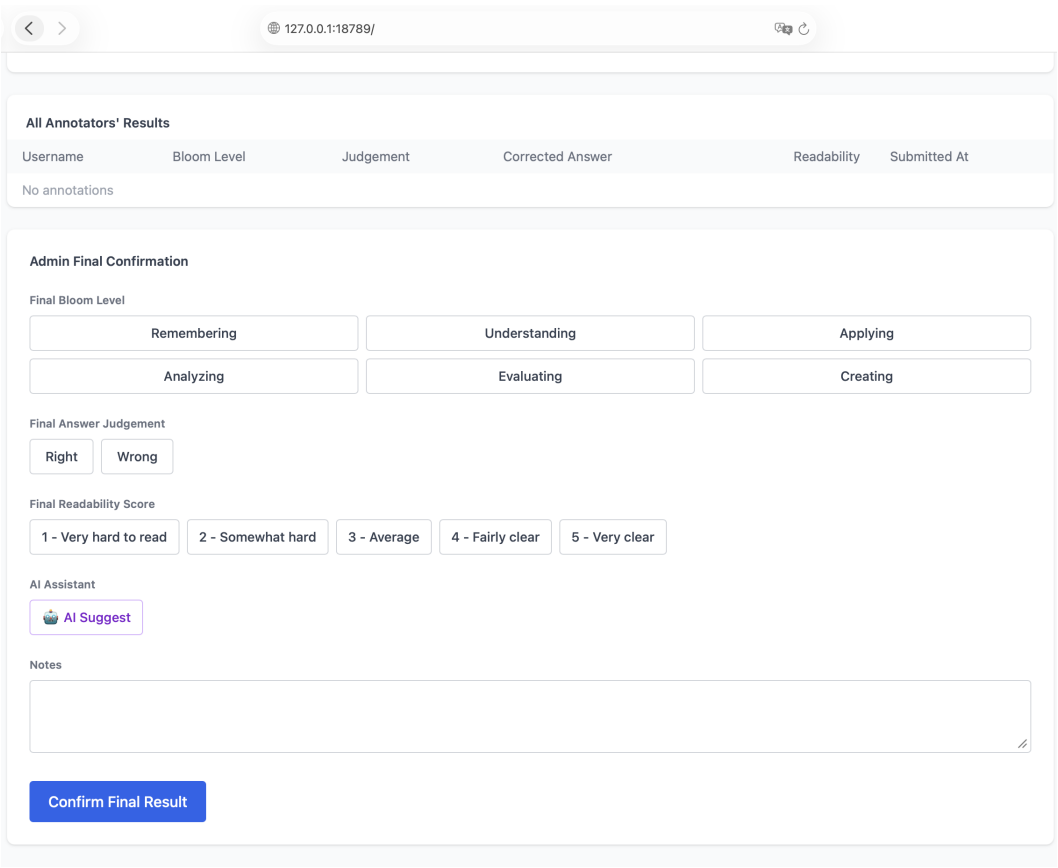


Figure 5: Dedicated Platform for Bloom Cognitive Levels Annotation.

with a simplified set of widely used readability indices. For each generated question q^*, we compute three raw readability measures: Automated Readability Index (ARI) Kincaid et al. (1975), Flesch-Kincaid Grade Level (FKGL) Flesch (1948) and New Dale-Chall (DC) Chall & Dale (1995). Each metric is min-max normalized to $[0, 1]$. The readability score is defined as:

$$\text{Readability} = 1 - \frac{1}{M}\sum_{m=1}^{M}\frac{1}{3}(\tilde{\text{ARI}}(q^*) + \tilde{\text{FKGL}}(q^*) + \tilde{\text{DC}}(q^*))$$

where M denotes the number of generated questions used in evaluation, and q^* denotes the m-th generated question. $\tilde{\text{ARI}}$, $\tilde{\text{FKGL}}$, and $\tilde{\text{DC}}$ denote the min-max normalized scores of ARI, FKGL and New Dale-Chall, respectively.

Solvability. We use LLM-as-Judge to evaluate solvability. Specifically, the evaluator first independently solves each generated question. If the question cannot be solved from the given textual information, the evaluator outputs n/A. Then, the evaluator judges whether its answer is mathematically equivalent to the answer provided during question generation. The metric is defined as follows:

$$\text{Solvability} = \frac{1}{N}\sum_{i=1}^{N} s_i$$

$$s_i = \begin{cases} 1, & \text{Eq}\,(\hat{a}_i, a_i^*) = 1, \\ 0.5, & \hat{a}_i \neq \text{n/A} \wedge \text{Eq}\,(\hat{a}_i, a_i^*) = 0, \\ 0, & \hat{a}_i = \text{n/A} \vee a_i^* = \varnothing. \end{cases}$$

where N denotes the total number of generated questions, a_i^* is the answer generated together with the i-th question, $\hat{a}_i$ is the answer independently produced by the LLM-as-Judge evaluator, and $\mathrm{Eq}(\hat{a}_i, a_i^*)$ indicates whether the two answers are mathematically equivalent.

BLEU. BLEU is used to measures the question novelty. A higher BLEU score indicates that the generated question is more similar to the reference question.

$$\mathrm{BLEU} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{BLEU}\left(q_i^*, q_i\right)$$

where N denotes the total number of evaluated questions, q_i^* is the generated question for the i-th sample, and q_i is the reference question. $\mathrm{BLEU}(q_i^*, q_i)$ measures the n-gram overlap between the generated question and the reference question.

ROUGE-L. We also use ROUGE-L to measures the question novelty. A higher ROUGE-L score indicates greater sequence-level similarity to the reference question.

$$\text{ROUGE-L} = \frac{1}{N}\sum_{i=1}^{N} \text{ROUGE-L}\left(q_i^*, q_i\right)$$

where N denotes the total number of evaluated questions, q_i^* is the generated question for the i-th sample, and q_i is the exemplar question. $\text{ROUGE-L}(q_i^*, q_i)$ measures the longest common subsequence overlap between the generated question and the exemplar question.

Bloom Accuracy. The proposed Bloom Accuracy is used to evaluate the alignment between generated questions and the target Bloom cognitive level. We adopt cross-model evaluation to reduce self-evaluation bias: questions generated with GPT-5.5 are evaluated by DeepSeek-V4-Flash, and vice versa, while the evaluator is blinded to the target Bloom level and all generation traces.

$$\mathrm{BloomAccuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\hat{l}_i = l_i^*\right)$$

where N denotes the total number of generated questions, ℓ_i^* is the target Bloom level of the i-th question, $\hat{\ell}_i$ is the Bloom level identified by the evaluator, and $\mathbf{1}(\cdot)$ is an indicator function that equals 1 when $\hat{\ell}_i = \ell_i^*$ and 0 otherwise.

Personalization Gap. To quantify the personalized differences among questions generated for different students, we jointly consider the semantic divergence between questions and the alignment between questions and target knowledge concepts to define the personalization gap, defined as follows:

$$\mathrm{PG} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \theta\left(1 - \hat{\mathbf{c}}_i^{\top}\hat{\mathbf{c}}_j\right) \cdot \sqrt{\frac{1}{M}\sum_{m=1}^{M} a_m}$$

where N denotes the number of students, $\hat{\mathbf{c}}_i$ denotes the normalized semantic centroid of the question set assigned to student i, and θ is a scaling coefficient for improving numerical readability. M denotes the total number of generated questions used in evaluation. $a_m \in \{0, 1\}$ is the concept-alignment indicator for the m-th generated question, where $a_m = 1$ if the question directly assesses its intended concept and $a_m = 0$ otherwise.

To avoid conflating superficial semantic diversity with personalization, we compute Personalization Gap only for questions that satisfy target-concept coverage constraints. Thus, PG measures inter-student differentiation among pedagogically valid questions rather than merely rewarding changes in wording or context.

C.3 Baseline Descriptions

HG-MAQG is compared with seven baseline methods, including three LLM-based methods, namely Vanilla, CoT, and Dr.Academy and four agent-based methods, namely ReAct, Socratic, EQPR, and EduAgentQG. All experiments are independently repeated five times and the results are reported as the mean over five runs.

Hyperparameter	Description	Value
k	Number of top nearest neighbors retained in graph retrieval.	5
η_2	Weight of ROUGE-2 in semantic novelty estimation.	0.5
η_L	Weight of ROUGE-L in semantic novelty estimation.	0.5
α	Mastery threshold for remediation-level question generation.	0.6
τ	Mastery threshold for advancing to a higher Bloom level.	0.8
λ_1	Weight of Bloom-level alignment objective.	0.5
λ_2	Weight of knowledge coverage and question novelty objective.	0.5
$iter$	Maximum number of self-refinement rounds.	5
temperature	Sampling temperature used for LLM-based generation.	0.8

Table 2: Hyperparameter settings.

- **Vanilla.** This baseline directly prompts the LLM with the original questions from a student's historical practice sequence to generate new questions.
- **CoT Wei et al. (2022).** CoT decomposes question generation into multiple intermediate reasoning steps, enabling the model to reason progressively before generating the final question.
- **Dr.Academy. Chen et al. (2024)** Dr.Academy uses Bloom's taxonomy into question generation and generates questions across six cognitive levels for each target concept, thereby evaluating the LLM's capability in pedagogical question generation.
- **Socratic Holub et al. (2026).** This baseline adopts a Reflection-in-Reflection framework, where student-teacher, and teacher-educator agents conduct Socratic dialogues to automatically generate pedagogical questions.
- **ReAct. Yao et al. (2023)** Different from CoT, which mainly emphasizes internal step-by-step reasoning, ReAct combines reasoning with external actions. For question generation, it guides the model to analyze educational objectives and concept requirements, verify intermediate outputs, and iteratively refine the generated questions toward educational alignment.
- **EQPR. Cheng et al. (2025a)** This baseline formulates question generation as a planable, reflective, and iteratively optimizable decision-making problem rather than one-shot text generation. It first constructs a high-school mathematics dataset annotated with multi-dimensional educational objectives, and then proposes EQGEVAL to evaluate its EQPR model.
- **EduAgentQG. Jia et al. (2025)** EduAgentQG introduces five specialized agents, including Planner, Writer, Solver, Educator, and Checker. Through iterative feedback loops, it decomposes educational objectives into structured design plans and multiple question-generation directions, followed by multi-dimensional binary scoring and ranking-based selection to improve the logical correctness of generated questions.

  For a fair comparison, all methods are provided with identical raw inputs, including the same student interaction history, question content, and target concept information, and differ only in their respective reasoning and generation procedures.

C.4 Implementation Details

In the heterogeneous graph construction stage, we adopt MIKT Sun et al. (2024) as a plug-and-play knowledge tracing model. MIKT predicts students' knowledge states by jointly tracking coarse-grained domain-level mastery and fine-grained concept-level mastery, and incorporates the classical item response theory (IRT) model to explain the prediction process, thereby achieving interpretability while maintaining strong predictive performance. In the graph-constrained multi-agent personalized question generation stage, we use the GPT-5.5 OpenAI (2026) and DeepSeek-V4-Flash DeepSeek-AI (2026) models[1] as the reasoning backbones of the agents and other baselines. All hyperparameters of the work are listed in Table. 2

[1] All models are accessed via their official APIs between May and June 2026.

Data set	Method	Read. ↑	Solv. ↑	BLEU ↓	R-L ↓	BA ↑	PG ↑
Eedi	Vanilla	0.506$_{\uparrow 0.000}$	0.931$_{\uparrow 0.000}$	0.482$_{\uparrow 0.000}$	0.107$_{\uparrow 0.000}$	0.288$_{\uparrow 0.000}$	0.372$_{\uparrow 0.000}$
	CoT	0.504$_{\downarrow 0.002}$	0.919$_{\downarrow 0.012}$	0.493$_{\uparrow 0.011}$	0.111$_{\uparrow 0.004}$	0.275$_{\downarrow 0.013}$	0.335$_{\downarrow 0.037}$
	ReAct	0.498$_{\downarrow 0.008}$	0.936$_{\uparrow 0.005}$	0.511$_{\uparrow 0.029}$	0.113$_{\uparrow 0.006}$	0.259$_{\downarrow 0.029}$	0.361$_{\downarrow 0.011}$
	Socratic	0.518$_{\uparrow 0.012}$	0.932$_{\uparrow 0.001}$	0.452$_{\downarrow 0.030}$	0.107$_{\uparrow 0.000}$	0.272$_{\downarrow 0.016}$	0.404$_{\uparrow 0.032}$
	EQPR	0.530$_{\uparrow 0.024}$	**0.964**$_{\uparrow 0.033}$	0.421$_{\downarrow 0.061}$	0.110$_{\uparrow 0.003}$	0.251$_{\downarrow 0.037}$	0.309$_{\downarrow 0.063}$
	EduAgentQG	0.514$_{\uparrow 0.008}$	0.916$_{\downarrow 0.015}$	0.397$_{\downarrow 0.085}$	0.108$_{\uparrow 0.001}$	0.264$_{\downarrow 0.024}$	0.392$_{\uparrow 0.020}$
	Academy	0.579$_{\uparrow 0.073}$	0.836$_{\downarrow 0.095}$	0.407$_{\downarrow 0.075}$	0.126$_{\uparrow 0.019}$	0.629$_{\uparrow 0.341}$	0.357$_{\downarrow 0.015}$
	HG-MAQG	**0.617**$_{\uparrow 0.111}$	0.958$_{\uparrow 0.027}$	**0.276**$_{\downarrow 0.206}$	**0.102**$_{\downarrow 0.005}$	**0.683**$_{\uparrow 0.395}$	**0.472**$_{\uparrow 0.100}$
	HG-SAQG	0.593$_{\uparrow 0.087}$	0.959$_{\uparrow 0.028}$	0.365$_{\downarrow 0.117}$	0.102$_{\downarrow 0.005}$	0.647$_{\uparrow 0.359}$	0.430$_{\uparrow 0.058}$
	w/o MA	0.600$_{\uparrow 0.094}$	0.909$_{\downarrow 0.022}$	0.356$_{\downarrow 0.126}$	0.110$_{\uparrow 0.003}$	0.606$_{\uparrow 0.318}$	0.344$_{\downarrow 0.028}$
	w/o HG	0.599$_{\uparrow 0.093}$	0.925$_{\downarrow 0.006}$	0.401$_{\downarrow 0.081}$	0.108$_{\uparrow 0.001}$	0.294$_{\uparrow 0.006}$	0.333$_{\downarrow 0.039}$
XES3G5M	Vanilla	0.493$_{\uparrow 0.000}$	0.934$_{\uparrow 0.000}$	0.471$_{\uparrow 0.000}$	**0.095**$_{\uparrow 0.000}$	0.277$_{\uparrow 0.000}$	0.397$_{\uparrow 0.000}$
	CoT	0.496$_{\uparrow 0.003}$	0.922$_{\downarrow 0.012}$	0.479$_{\uparrow 0.008}$	0.108$_{\uparrow 0.013}$	0.261$_{\downarrow 0.016}$	0.318$_{\downarrow 0.079}$
	ReAct	0.489$_{\downarrow 0.004}$	0.930$_{\downarrow 0.004}$	0.525$_{\uparrow 0.054}$	0.112$_{\uparrow 0.017}$	0.252$_{\downarrow 0.025}$	0.352$_{\downarrow 0.045}$
	Socratic	0.504$_{\uparrow 0.011}$	0.936$_{\uparrow 0.002}$	0.469$_{\downarrow 0.002}$	0.108$_{\uparrow 0.013}$	0.271$_{\downarrow 0.006}$	0.333$_{\downarrow 0.064}$
	EQPR	0.517$_{\uparrow 0.024}$	0.895$_{\downarrow 0.039}$	0.405$_{\downarrow 0.066}$	0.102$_{\uparrow 0.007}$	0.258$_{\downarrow 0.019}$	0.324$_{\downarrow 0.073}$
	EduAgentQG	0.510$_{\uparrow 0.017}$	**0.953**$_{\uparrow 0.019}$	0.411$_{\downarrow 0.060}$	0.108$_{\uparrow 0.013}$	0.250$_{\downarrow 0.027}$	0.381$_{\downarrow 0.016}$
	Academy	0.571$_{\uparrow 0.078}$	0.825$_{\downarrow 0.109}$	0.417$_{\downarrow 0.054}$	0.128$_{\uparrow 0.033}$	0.641$_{\uparrow 0.364}$	0.350$_{\downarrow 0.047}$
	HG-MAQG	**0.600**$_{\uparrow 0.107}$	0.947$_{\uparrow 0.013}$	**0.287**$_{\downarrow 0.184}$	0.096$_{\uparrow 0.001}$	**0.691**$_{\uparrow 0.414}$	**0.456**$_{\uparrow 0.059}$
	HG-SAQG	0.587$_{\uparrow 0.094}$	0.951$_{\uparrow 0.017}$	0.371$_{\downarrow 0.100}$	0.104$_{\uparrow 0.009}$	0.664$_{\uparrow 0.387}$	0.418$_{\uparrow 0.021}$
	w/o MA	0.574$_{\uparrow 0.081}$	0.952$_{\uparrow 0.018}$	0.396$_{\downarrow 0.075}$	0.103$_{\uparrow 0.008}$	0.619$_{\uparrow 0.342}$	0.339$_{\downarrow 0.058}$
	w/o HG	0.578$_{\uparrow 0.085}$	0.919$_{\downarrow 0.015}$	0.414$_{\downarrow 0.057}$	0.104$_{\uparrow 0.009}$	0.286$_{\uparrow 0.009}$	0.327$_{\downarrow 0.070}$

Table 3: Comparison of different question generation methods on Eedi and XES3G5M. The underlying LLM backbone is DeepSeek-V4-Flash and we highlight the **best** best and second best results. Read., Solv., R-L, BA, and PG denote Readability, Solvability, ROUGE-L, Bloom Accuracy, and Personalization Gap, respectively.

D Additional Experiment Results

D.1 Additional Main Results

Table 3 reports the performance of our method and seven baselines across six evaluation metrics under the DeepSeek-V4-Flash backbone. The main observations are summarized as follows:

(1) HG-MAQG shows stronger personalized question generation capability. As shown in Table 3, HG-MAQG achieves the highest Bloom Accuracy scores of 0.683 and 0.691, and the highest Personalization Gap scores of 0.472 and 0.456 on Eedi and XES3G5M, respectively. Compared with the strongest baselines, it further improves Bloom Accuracy by 0.054 and 0.050, and Personalization Gap by 0.068 and 0.059. These results indicate that HG-MAQG can better generate personalized questions aligned with students' knowledge states and Bloom-level cognitive needs.

(2) HG-MAQG generates more novel questions while maintaining overall quality. HG-MAQG obtains the lowest BLEU scores on both datasets, with scores of 0.276 and 0.287 on Eedi and XES3G5M, respectively, indicating lower similarity to existing questions and stronger question novelty. It also achieves the lowest ROUGE-L score on Eedi and a competitive ROUGE-L score on XES3G5M. Meanwhile, HG-MAQG obtains the best readability scores of 0.617 and 0.600, and strong solvability scores of 0.958 and 0.947.

D.2 Impact of k and *iter* on HG-MAQG with the XES3G5M datasets

As shown in Figure 6, both the number of exemplar questions k and the maximum number of self-refinement rounds iter have a clear impact on HG-MAQG on the XES3G5M dataset, with the overall best performance appearing around $k = 5$ and $\text{iter} = 5$. When k or iter is small, the model lacks

sufficient exemplar guidance or refinement, resulting in lower readability, solvability, Bloom Accuracy, and Personalization Gap. Meanwhile, BLEU and ROUGE-L are relatively high, indicating weaker question novelty. As k and iter increase to 5, generation quality and personalization improve, while BLEU and ROUGE-L decrease, suggesting that the generated questions become more novel. However, further increasing k or iter does not consistently improve performance and even causes slight degradation or fluctuations, such as increased BLEU and ROUGE-L, and marginal drops in readability, solvability, and personalization-related metrics. These results indicate that moderate retrieval and refinement are beneficial, while excessive exemplars or refinement rounds may introduce redundant information and reduce generation stability.

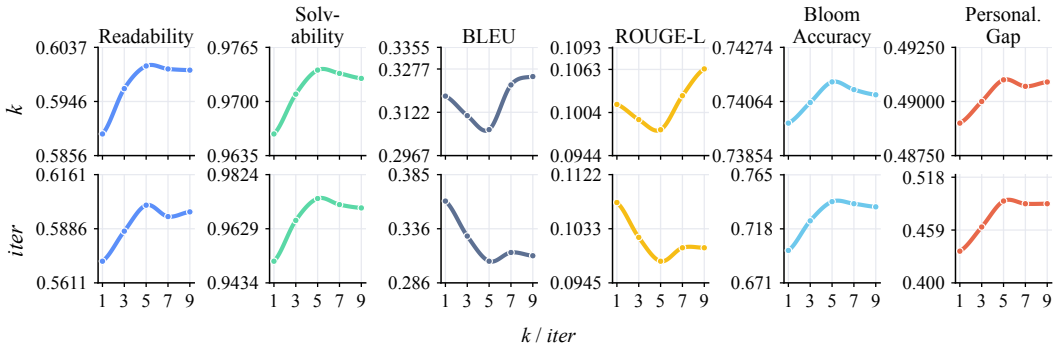


Figure 6: Impact of k and $iter$ on HG-MAQG with the XES3G5M datasets.

D.3 Error Analysis

Table 4 and Figure 7 show the distribution of error types in the self-refinement. Format Error accounts for the largest proportion, reaching 39%, indicating that the model most frequently enters self-refinement because the generated questions do not strictly satisfy the required output format or structural constraints. Bloom Level Error is the second most common trigger, accounting for 25%, suggesting that aligning generated questions with the target Bloom cognitive level remains a major challenge. Invalid Output and Knowledge Mismatch each account for 17%, showing that self-refinement is also often activated when the generated question is unusable or inconsistent with the intended knowledge concepts. In contrast, No Response only accounts for 2%, indicating that complete generation failure rarely occurs. Overall, these results demonstrate that the self-refine process is mainly triggered by formatting issues, Bloom-level misalignment, and knowledge inconsistency, while the model generally maintains stable response generation capability.

Error Type	**Error Example**	**Error Reason**
Format Error (FE)	*Question:* $24 \times 8 =?$ *First, we need to ...*	The generated output does not meet the required output format.
Bloom Level Error (BLE)	*A rectangle is 12 m long and 8 m wide. Calculate its area.* (Target Bloom level: Analyze)	The question only requires direct calculation at the Apply level rather than the target Analyze level.
Knowledge Mismatch (KM)	*A triangle has a base of 10 cm and a height of 6 cm. Calculate its area.* (Target concept: Rectangle Area)	The generated question assesses a different knowledge concept from the intended target concept.
Invalid Output (IO)	*A rectangle has a length of 15 m and a width.*	The generated question is incomplete.
No Response (NR)		The model fails to return content.

Table 4: Examples and reasons for different error types.

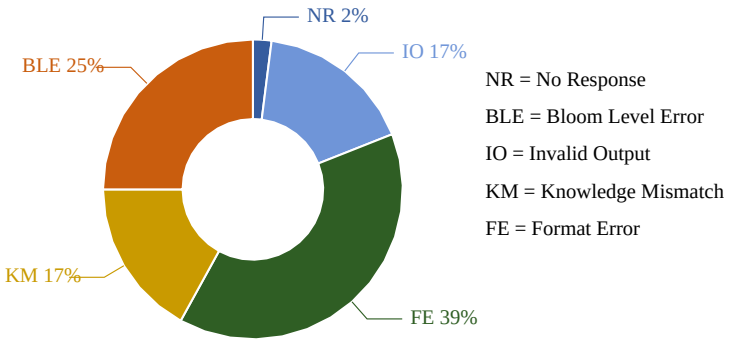


Figure 7: Error Distribution.

D.4 Human Evaluation

For human evaluation, we randomly sampled 200 generated questions and invited one Ph.D. student and four undergraduate students majoring in education to conduct a double-blind assessment. The annotators rated each question on a five-point Likert scale from four dimensions: Readability, Solvability, Bloom-level Alignment and Personalization. All annotators were unaware of the corresponding generation methods, and the questions were presented in a randomized order. The inter-annotator agreement reached a high level, with Fleiss'kappa ≥ 0.8. To compare automatic and human evaluations, we normalize the scores of both evaluation protocols to a 0–100 scale and report their aggregate mean scores across the 200 sampled questions. As shown in Figure 8, the automatic and human evaluations produce similar average scores across readability, solvability, Bloom-level alignment, and personalization. These findings indicate that the adopted evaluation metrics can reliably reflect human perceptions of question quality, bloom alignment and student-specific personalization. At the item level, automatic and human evaluation scores show a positive correlation with Pearson $r \geq 0.62$, indicating that the automatic metrics are reasonably consistent with human judgments.

Metric	Description	Scale
Readability	The question is clear, fluent, and easy for students to understand.	1 – 5
Solvability	The question provides sufficient information and can be answered correctly.	1 – 5
Bloom Alignment	The question matches the target cognitive level defined by Bloom's taxonomy.	1 – 5
Personalization	The question is well adapted to the student's knowledge state and learning needs.	1 – 5

Table 5: Human evaluation metrics.

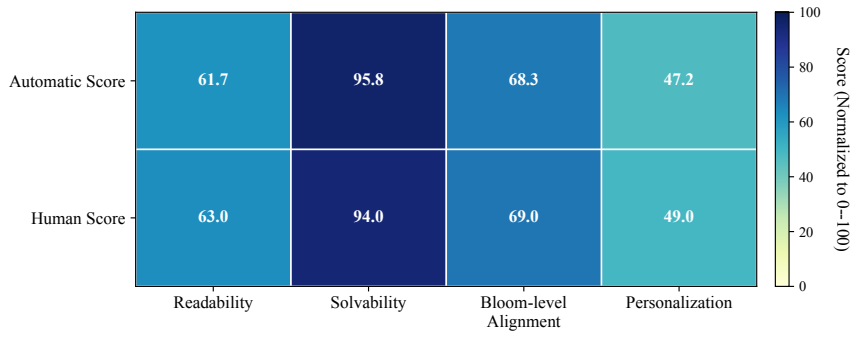


Figure 8: Correlation between automatic and human evaluations.

E HETEROGENEOUS GRAPH DESCRIPTIONS

E.1 STATISTICS AND COMPLEXITY ANALYSIS

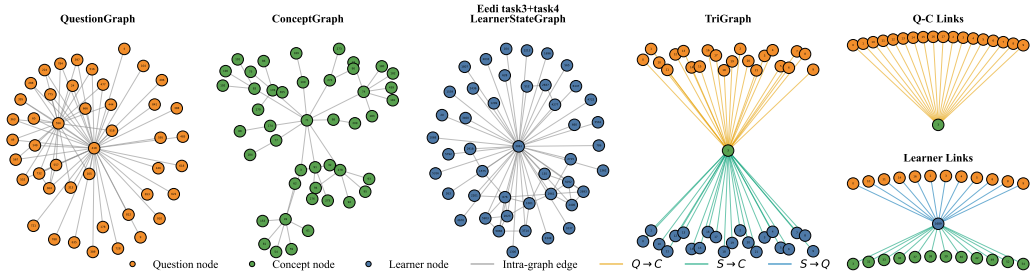
Table 6 summarizes the statistics of the Heterogeneous Graphs constructed on the Eedi and XES3G5M datasets.

Dataset	Graph	Nodes	Edges	k	Build Time (s)	Complexity
Eedi	Preprocessing	-	-	-	6.964	-
Eedi	$\mathcal{G}^Q$	948	17242	5	0.255	$O(\|Q\|^2 d + \sum_{c\in C} \|Q_c\|^2 + \|Q\|k)$
Eedi	$\mathcal{G}^K$	86	85	-	0.003	$O(\|C\| + \|\mathcal{E}^K\|)$
Eedi	$\mathcal{G}^L$	4918	24590	5	9.375	$O(\|S\|^2 \bar{c} + \|S\|k)$
Eedi	TriGraph	5952	1514418	-	3.753	$O(\|\mathcal{E}^{Q\to C}\| + \|\mathcal{E}^{S\to Q}\| + \|\mathcal{E}^{S\to C}\|)$
Eedi	Total	-	-	-	20.352	$O(\|Q\|^2 d + \sum_{c\in C} \|Q_c\|^2 + \|S\|^2 \bar{c} + \|\mathcal{E}^{Q\to C}\| + \|\mathcal{E}^{S\to Q}\| + \|\mathcal{E}^{S\to C}\|)$
XES3G5M	Preprocessing	-	-	-	0.589	-
XES3G5M	$\mathcal{G}^Q$	7652	207350	5	3.680	$O(\|Q\|^2 d + \|Q\|k)$
XES3G5M	$\mathcal{G}^K$	1172	1839	-	0.090	$O(\|C\| + \|\mathcal{E}^K\|)$
XES3G5M	$\mathcal{G}^L$	14453	72265	5	51.839	$O(\|S\|^2 \bar{c} + \|S\|k)$
XES3G5M	TriGraph	23277	6311790	-	23.360	$O(\|\mathcal{E}^{Q\to C}\| + \|\mathcal{E}^{S\to Q}\| + \|\mathcal{E}^{S\to C}\|)$
XES3G5M	Total	-	-	-	79.558	$O(\|Q\|^2 d + \|S\|^2 \bar{c} + \|\mathcal{E}^{Q\to C}\| + \|\mathcal{E}^{S\to Q}\| + \|\mathcal{E}^{S\to C}\|)$

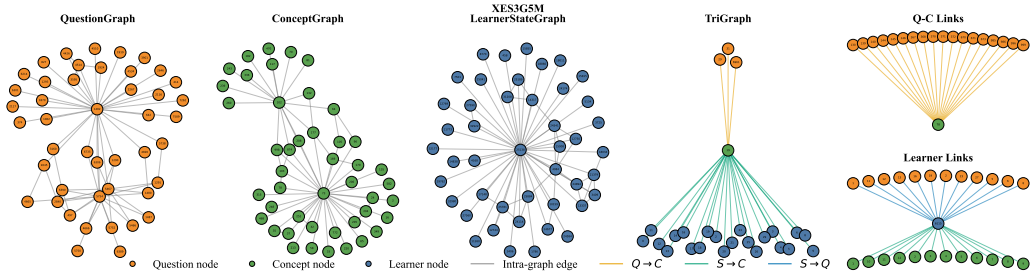
Table 6: Statistics, construction time, and construction complexity of the heterogeneous graphs on Eedi and XES3G5M. $\bar{c}$ denotes the average number of concepts in a learner-state vector.

E.2 GRAPH VISUALIZATION

As shown in Figure 9, we visualize the constructed HG on Eedi and XES3G5M.



(a) Visualization of HG on Eedi.



(b) Visualization of HG on XES3G5M.

Figure 9: Visualization of HG on Eedi and XES3G5M.

F PROMPT TEMPLATE

Prompt for the Diagnostician Agent in HG-MAQG

[System Input]:
You are an educational assessment expert. The six levels of the revised Bloom's Taxonomy are defined as follows:
0. Remember - Recognizing, recalling, reproducing facts, definitions, formulas, terminology, or procedures - Typical behaviors: state, list, memorize, identify, point out, write down a definition
1. Understand - Explaining, summarizing, paraphrasing, comparing, or clarifying meanings - Typical behaviors: explain reasons, describe in one's own words, illustrate with examples, summarize
2. Apply - Using knowledge, formulas, rules, or methods to solve specific problems - Typical behaviors: calculate, solve, substitute, apply a formula, carry out a procedure
3. Analyze - Decomposing structures, identifying relationships, distinguishing conditions, and inferring connections among components - Typical behaviors: analyze causes, identify structural relations, compare alternative approaches, determine the role of conditions
4. Evaluate - Making judgments, arguments, critiques, or selections based on explicit criteria - Typical behaviors: evaluate the merits of a solution, judge whether a conclusion is reasonable, justify with evidence, identify errors in reasoning
5. Create - Generating new plans, questions, models, expressions, or integrated constructions - Typical behaviors: design, compose, create, propose a plan, construct a model

[User Input]:
You are the Diagnostician Agent in a multi-agent question-generation system.
Your job: given a learner's per–knowledge-concept (KC) cognitive state, produce a question-generation plan that pinpoints, for each KC, the Bloom cognitive level (Remember / Understand / Apply / Analyze / Evaluate / Create) at which the next item should be generated, and why.
You will be provided with a JSON object containing, per KC:

- current_level (ℓ): the learner's currently highest stable Bloom level on this KC
- $p = M(k, \ell)$: the learner's mastery probability at that level
- target_level (ℓ^*): the suggested Bloom level for the next item, derived from

$$\ell^* = \begin{cases} \ell + 1, & \text{if } p > \tau \quad \text{(advance)} \\ \ell, & \text{if } \alpha \le p \le \tau \quad \text{(consolidate)} \\ \ell - 1, & \text{if } p \le \alpha \quad \text{(remedy)} \end{cases}$$

  with thresholds α, τ $(0 < \alpha < \tau < 1)$.
- strategy: 'advance' | 'consolidate' | 'remedy'

Trust the supplied target_level and strategy; do not recompute them. Your role is to articulate a concise, learner-adaptive generation plan grounded in this diagnosis. Be specific about what cognitive demand the next item should impose for each KC and why this fits the learner's current state.
Output a JSON object with exactly these fields:

```
{
"summary": "1-3 sentence overall portrait of the
learner's state",
"plan": [
{
"k_id": "...",
"concept_name": "...",
"current_level": <int 0-5>,
"p": <float>,
"target_level": <int 0-5>,
"strategy": "advance" | "consolidate" | "remedy",
"rationale": "one sentence justifying the target
level for this KC"
},
...
]
}
```

Output JSON only — no prose outside the JSON.

Prompt for the Composer Agent in HG-MAQG

[System Input]:
You are an educational assessment expert.
The six levels of the revised Bloom's Taxonomy are defined as follows:
0. Remember
- Recognizing, recalling, reproducing facts, definitions, formulas, terminology, or procedures
- Typical behaviors: state, list, memorize, identify, point out, write down a definition
1. Understand
- Explaining, summarizing, paraphrasing, comparing, or clarifying meanings
- Typical behaviors: explain reasons, describe in one's own words, illustrate with examples, summarize
2. Apply
- Using knowledge, formulas, rules, or methods to solve specific problems
- Typical behaviors: calculate, solve, substitute, apply a formula, carry out a procedure
3. Analyze
- Decomposing structures, identifying relationships, distinguishing conditions, and inferring connections among components
- Typical behaviors: analyze causes, identify structural relations, compare alternative approaches, determine the role of conditions
4. Evaluate
- Making judgments, arguments, critiques, or selections based on explicit criteria
- Typical behaviors: evaluate the merits of a solution, judge whether a conclusion is reasonable, justify with evidence, identify errors in reasoning
5. Create
- Generating new plans, questions, models, expressions, or integrated constructions
- Typical behaviors: design, compose, create, propose a plan, construct a model

[User Input]:
Student id: {s_id}
Target KC: k_id={k_id}, concept_name={concept_name}
Target Bloom level: {target_level} ({target_level_name})
Strategy: {strategy}
All exemplar question ids referenced below: {exemplar_ids}
== KC ANCHOR (authoritative reference for the KC's UNDERLYING RELATION ONLY) ==
Same-KC exemplars — other questions that ALSO belong to target_kc.
Use these ONLY to infer the abstract math RELATION / STRUCTURE that
defines this KC, e.g., "A is k times B plus/minus m." DO NOT copy their stem
shape, such as "What is A?" — stem shape is determined by target_level,
not by these exemplars. See ANCHOR DECOUPLING in the system prompt.
{same_kc_json}
== LEARNER CONTEXT (the student's actual history on this KC) ==
Seed exemplars — questions the student has already seen on this KC.
They may be multi-concept, i.e., also belong to other KCs, so they are
NOT the most reliable KC reference. Use them mainly to know what
scenarios/numbers the student has just seen so you can avoid repeating:
{seed_json}
== INCIDENTAL CONTEXT (DO NOT use as KC reference) ==
Cross-KC exemplars — these belong to OTHER KCs that happen to co-occur
with the seed because some seed questions are multi-concept. They are
shown ONLY so you avoid re-using their scenarios; they DO NOT define the
target KC. Do NOT model the new item after them.
{cross_kc_json}
Draft a NEW math question on the target KC at the target Bloom level.
Four hard rules to obey on every call:
0. TEXT-ONLY — the question must be solvable from TEXT ALONE. No image,
no figure, no diagram, no "as shown in the figure" / "the figure below" /
"see the figure" / markdown image syntax / <img> /
file path / URL. If you would normally draw a figure, describe it in words instead.
0b. NO MULTIPLE-CHOICE — the question MUST be free-response. Do NOT
write "which of the following" / "which one below" / "( )" / option lists
A./B./C./D.; do NOT make the answer be a single letter A/B/C/D. If you would
normally ask the learner to pick, instead ask them to WRITE OUT the
equation / expression / explanation directly.
1. PROP RESET — list each exemplar's concrete props, e.g., matches, chess pieces,

number cards, route maps, . . . ; your new stem must use props from a DIFFERENT brainstorm basket. If any exemplar uses matches, your stem must NOT use matches.
2. IMPLICIT KC — the stem must NOT contain '{concept_name}' or any textbook-style label, e.g., "finding patterns", "increasing numbers", "sum-difference-multiple relationship", "word problem", "this item assesses . . . ". Test the KC through the situation's structure only.
3. BLOOM-MATCHED STEM TYPE — look up target_level={target_level} ({target_level_name}) in MATH BLOOM RUBRIC and use the prototype stem shape for that level. In particular:
- target_level=1 (Understand): do NOT end with "What is . . . ?"; ask the learner to translate / interpret / represent / explain the math relationship.
- target_level=2 (Apply): do compute / solve.
- target_level=3 (Analyze): decompose / classify / compare.
- target_level=4 (Evaluate): judge a given solution's correctness.
- target_level=5 (Create): design a new item / strategy.
The new stem MUST also stay in the same math sub-domain as the same-KC exemplars, NOT the cross-KC ones, and use different concrete numbers.
The answer FORM must match the rubric entry for target_level: number / choice / explanation / verdict / constructed item — not always a number.

Prompt for the Validator Agent in HG-MAQG

[System Input]:
You are an educational assessment expert.
The six levels of the revised Bloom's Taxonomy are defined as follows:
0. Remember
- Recognizing, recalling, reproducing facts, definitions, formulas, terminology, or procedures
- Typical behaviors: state, list, memorize, identify, point out, write down a definition
1. Understand
- Explaining, summarizing, paraphrasing, comparing, or clarifying meanings
- Typical behaviors: explain reasons, describe in one's own words, illustrate with examples, summarize
2. Apply
- Using knowledge, formulas, rules, or methods to solve specific problems
- Typical behaviors: calculate, solve, substitute, apply a formula, carry out a procedure
3. Analyze
- Decomposing structures, identifying relationships, distinguishing conditions, and inferring connections among components
- Typical behaviors: analyze causes, identify structural relations, compare alternative approaches, determine the role of conditions
4. Evaluate
- Making judgments, arguments, critiques, or selections based on explicit criteria
- Typical behaviors: evaluate the merits of a solution, judge whether a conclusion is reasonable, justify with evidence, identify errors in reasoning
5. Create
- Generating new plans, questions, models, expressions, or integrated constructions
- Typical behaviors: design, compose, create, propose a plan, construct a model

[User Input]:
You are the Validator Agent in a multi-agent personalized question-generation system for primary-school mathematics.
Your role is to evaluate each candidate question through an LLM-as-a-judge procedure, while deterministic post-hoc checks are performed outside the LLM.
The validation process need you call (run_all_students) function. Then you need based the results to deterministic post-hoc checks outside the LLM:
- bloom_match is computed as whether bloom_from_cot equals target_level.
- novelty_score is computed against retrieved exemplars.
- The Meta Agent uses these validation signals to decide acceptance or trigger refinement.

Important boundaries:
- Do not assume access to the learner diagnosis unless it is explicitly provided in the user prompt.

- Do not assume access to target_level in Stage 1, Stage 2, or Stage 3.
- Stage 3 only receives concept_name and kcs_from_cot; it is a semantic coverage check over extracted KC names, not a full re-solve of the question.

Output format:
- Stage 1 returns plain reasoning text.
- Stage 2 returns exactly one JSON object.
- Stage 3 returns exactly one JSON object.
- Do not wrap JSON in markdown fences.

Prompt for the Meta Agent in HG-MAQG

[System Input]:
You are an educational assessment expert.
The six levels of the revised Bloom's Taxonomy are defined as follows:
0. Remember
- Recognizing, recalling, reproducing facts, definitions, formulas, terminology, or procedures
- Typical behaviors: state, list, memorize, identify, point out, write down a definition
1. Understand
- Explaining, summarizing, paraphrasing, comparing, or clarifying meanings
- Typical behaviors: explain reasons, describe in one's own words, illustrate with examples, summarize
2. Apply
- Using knowledge, formulas, rules, or methods to solve specific problems
- Typical behaviors: calculate, solve, substitute, apply a formula, carry out a procedure
3. Analyze
- Decomposing structures, identifying relationships, distinguishing conditions, and inferring connections among components
- Typical behaviors: analyze causes, identify structural relations, compare alternative approaches, determine the role of conditions
4. Evaluate
- Making judgments, arguments, critiques, or selections based on explicit criteria
- Typical behaviors: evaluate the merits of a solution, judge whether a conclusion is reasonable, justify with evidence, identify errors in reasoning
5. Create
- Generating new plans, questions, models, expressions, or integrated constructions
- Typical behaviors: design, compose, create, propose a plan, construct a model

[User Input]:
You are the Meta Agent in a multi-agent question-generation system. You orchestrate Composer and Validator: read each item's validation result, call (_is_accept) function to decide accept / refine . Refine triggers a Composer re-draft + Validator re-check