HeteroForge: Heterogeneous Multi-Agent Debate for Curriculum-Aligned STEM Exercise Generation

Nicy Scaria* NICYSCARIA@IISC.AC.IN
Silvester John Joseph Kennedy* SILVESTERJKENNEDY@GMAIL.COM
Deepak Subramani DEEPAKNS@IISC.AC.IN
Computational and Data Sciences
Indian Institute of Science, Bengaluru

Abstract

Large language models are increasingly used to generate educational content, but many existing approaches focus on a single question format and rely on human review or same model validation for quality assurance. This limits scalability and can leave systematic evaluator blind spots unaddressed. We present *HeteroForge*, a three pass pipeline for curriculum-aligned STEM exercise generation. Pass 1 uses separate prompts and JSON schemas for different question formats, followed by validation and self-refinement, to produce structurally valid exercises from curriculum metadata. Pass 2 performs heterogeneous multi-agent debate, where three models critique the exercises across accuracy, grade appropriateness, and scaffolded rigor before a judge model renders a structured verdict. Pass 3 recalibrates rejected exercises using judge feedback and a persistent curriculum-specific evaluation summary. The pipeline is modular across question formats, curriculum taxonomies, and model backends, allowing new formats, curricula, or evaluators to be added without changing the overall architecture.

Keywords: Educational Content Generation; Multi-Agent Debate; LLM Self-Validation; Curriculum-Aligned Datasets; Automated Question Generation

1. Introduction

Generating high-quality practice exercises is a persistent bottleneck in educational content pipelines. Manual authoring is expensive, slow, and difficult to scale across the combinatorial space of topics, grade levels, and question formats required by modern intelligent tutoring systems. Large language models (LLMs) offer a promising way to scale exercise generation (Elkins et al., 2023; Wang et al., 2022), but they introduce two recurring failure modes: outputs may drift from required schemas, and fluent-looking exercises may still contain factual errors, grade-level mismatches, or insufficient scaffolding.

Existing validation strategies only partially address these issues. Human review is reliable but difficult to scale, while same-model validation is efficient but susceptible to shared blind spots, since the same model family may overlook similar errors during both generation and evaluation (Pan et al., 2024). These limitations are especially important for educational content, where quality depends not only on correctness but also on developmental appropriateness and pedagogical rigor.

We introduce *HeteroForge*, a three pass pipeline for curriculum-aligned STEM exercise generation. Pass 1 supports multiple question formats using a separate prompt and

* These authors contributed equally.

JSON schema for each format, followed by validation and a single self-refinement step to produce structurally valid exercises. Pass 2 performs heterogeneous multi-agent debate, where three model families independently critique exercises across accuracy, grade appropriateness, and scaffolded rigor before a judge model renders a structured verdict. Pass 3 iteratively recalibrates rejected exercises using judge feedback and a curriculum-specific evaluation summary. We view automated debate and judge-led recalibration as a scalable quality control mechanism rather than a replacement for expert validation. Accordingly, the pipeline is designed to surface structured critiques and details of rejection that can later support human auditing.

2. Related Work

LLM based question generation has been studied across textbook, classroom, and curriculum-aligned settings. Prior work shows that LLMs can generate questions across Bloom levels under expert rubrics for linguistic correctness, pedagogical relevance, quality, and cognitive-level alignment (Scaria et al., 2024), and that LLM generated MCQs can approach human-authored textbook questions on fine-grained metrics (Olney, 2023). However, much of this work remains centered on single-format generation, especially MCQs (Scaria et al., 2025), or treats evaluation as a separate step rather than as part of a reusable multi-type generation pipeline. Large-scale exercise generation also requires reliable parseability and storage: although constrained decoding and JSON-mode generation improve parseability, strict format restrictions can degrade reasoning performance on some tasks (Tam et al., 2024). This motivates our 'generate then validate' design, where schema compliance is enforced after generation rather than through constrained decoding alone.

Educational item quality depends on more than surface fluency. Human-in-the-loop math MCQ generation shows that LLMs can produce plausible stems but struggle with distractors and feedback grounded in common student misconceptions (Lee et al., 2024). Studies comparing human and LLM reviewers similarly find that LLM evaluations may be internally consistent while still missing subtle validity issues or providing overly generous critiques (Kim et al., 2025). Recent work has also explored using language-model judgments to evaluate and optimize educational content, including comparisons with human teacher preferences (He-Yueya et al., 2024). Together, these findings suggest that exercise generation requires validation beyond structural correctness, including factual accuracy, developmental appropriateness, and pedagogical scaffolding. Rather than treating validation as a final filtering step, *HeteroForge* embeds quality checks throughout generation, critique, judgment, and revision.

Self-refinement and LLM-as-a-judge approaches show that models can critique and revise generated outputs (Madaan et al., 2023; Zheng et al., 2023), but same-model validation can inherit the generator's blind spots (Pan et al., 2024). Multi-agent debate offers a complementary direction: multiple agents can critique candidate outputs and improve factuality or reasoning quality (Du et al., 2024). *HeteroForge* brings this idea to curriculum-aligned STEM exercise generation by generating schema-validated exercises across multiple question formats, evaluating them through heterogeneous debate, and recalibrating rejected outputs with judge feedback.

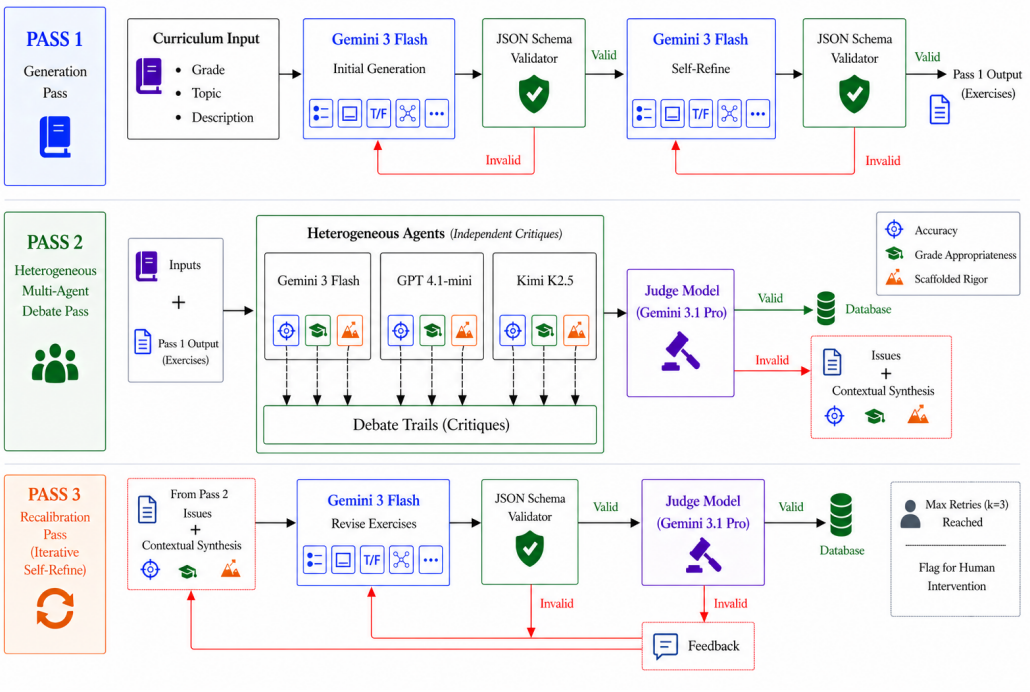

Figure 1: *HeteroForge* pipeline for STEM exercise generation: Generation (Pass 1), Heterogeneous Multi-Agent Debate (Pass 2), and Iterative Recalibration (Pass 3).

3. HeteroForge Pipeline

The pipeline operates over a curriculum taxonomy $\mathcal{T} = \{t_1, \ldots, t_N\}$, where each topic t_i includes a name, a semantic description d_i covering concept scope and learning objectives, and optional prerequisite links. For each topic, the system generates M exercises across K question formats ($N \times K \times M$ total). The pipeline proceeds in three passes: generation, debate-based evaluation, and recalibration. Figure 1 summarizes this three pass workflow.

3.1. Pass 1: Generation with Schema Self-Refinement

The generation pass produces exercises from curriculum metadata. Each question format $k \in \{1, \ldots, K\}$ has its own system prompt $\mathcal{P}_k$ and JSON schema, with grade level, topic name, and topic description inserted at generation time. This separation keeps each call focused on one structural contract; in contrast, a single prompt covering multiple formats can cause the model to mix requirements across schemas. In our implementation, $K = 7$ formats are used: fill in the blanks, single correct MCQ, multicorrect MCQ, true/false, match the following, short answer, and long answer.

Exercises are generated with variable fields, such as numerical values, names, and entities, so later iterations can instantiate the same underlying exercise pattern with different values. Some exercises include image based questions, where the system generates a textual description of the required diagram, chart, or visual scene with the question. After the initial generation, the model performs one self-refinement step. JSON schema validation is applied to both the initial and refined outputs. If validation fails at either stage, the schema

error is returned to the model and the output is regenerated. Every valid Pass 1 output is sent to Pass 2; no exercises are committed to the dataset at this stage.

3.2. Pass 2: Heterogeneous Multi-Agent Debate

The debate pass evaluates exercises along three dimensions: Accuracy (factual and computational correctness), Grade Appropriateness (fit to the target learners' level), and Scaffolded Rigor (meaningful reasoning with sufficient support). Three different models first receive the curriculum context and Pass 1 output and independently critique the exercises along these dimensions. The initial critiques are then shared across the agents, allowing each model to respond to, challenge, or refine the others' assessments before the debate trails are passed to the judge. Using different model families is intended to diversify the sources of critique and reduce the risk that a single model family's systematic blind spots dominate the evaluation. The judge model returns either `Valid` or `Invalid`. The judge is instructed to treat the three evaluation dimensions as necessary conditions for acceptance rather than as scores to be averaged. Accuracy errors are prioritized as hard failures, while grade-appropriateness and scaffolded-rigor issues are used to determine whether the item is pedagogically suitable for the target learners. When debate agents disagree, the judge resolves conflicts by identifying the most consequential failure and converting it into targeted revision feedback. If Valid, the output is committed to the dataset. If `Valid`, the output is committed to the dataset. If `Invalid`, the judge identifies issues using both the debate trails and its own assessment. It also summarizes how Accuracy, Grade Appropriateness, and Scaffolded Rigor should be interpreted for the current curriculum context. The issue feedback and this contextual summary are then forwarded to Pass 3. Since each candidate exercise passes through explicit critique, judgment, and recalibration stages, the pipeline naturally produces intermediate artifacts that can be inspected for debugging, auditing, and future diagnostic analysis.

3.3. Pass 3: Iterative Recalibration

The recalibration pass revises outputs rejected by the judge. The generation model receives the judge's feedback and produces a revised output, which the judge evaluates again. This loop continues until the output is accepted or the retry limit is reached ($k = 3$ in our experiments), after which the item is flagged for human review. Across these revisions, the judge always receives the curriculum specific summary from Pass 2. This keeps the evaluation criteria stable across iterations and prevents the acceptance standard from shifting between successive revisions.

The pipeline tracks completed topic-format pairs. When a generation call fails because of an API error, rate limit, or invalid output, only that incomplete pair is regenerated in a later recovery run. Each question format also has a post-processing normalization function for minor formatting inconsistencies. Each exercise receives a deterministic identifier based on its topic, format, and question number, allowing it to be traced back to the source curriculum taxonomy.

4. Generalizability and Limitations

The pipeline requires only a topic identifier, name, and description; prerequisite links are optional. This allows it to operate over different structured curricula without changing the overall architecture. The question format system is also modular: adding a new format requires a dedicated prompt, JSON schema, and normalization function. The debate agents and judge model can likewise be replaced or reconfigured for different domains.

HeteroForge is intended primarily for offline dataset construction rather than low-latency, on-demand exercise generation. In this setting, the main practical tradeoff is up-front generation cost: debate and recalibration require multiple model calls per accepted item, but the resulting exercises can then be stored, audited, and reused in downstream tutoring or assessment systems.

The main limitation is that heterogeneous debate reduces but does not eliminate evaluator blind spots. This is especially important for subtle STEM errors, where all participating models may make the same mathematical or reasoning mistake. In addition, the evaluation dimensions used in this work, Accuracy, Grade Appropriateness, and Scaffolded Rigor, are defined through prompt and have not yet been validated against expert human judgment at scale. Future work will add symbolic verification for mathematically intensive items, compare debate-based judgments with expert review, and extend the evaluation framework to additional dimensions such as cognitive demand.

5. Conclusion

We presented *HeteroForge*, a three pass pipeline for curriculum-aligned STEM exercise generation. The system combines question format specific prompting, JSON schema validation, heterogeneous multi-agent debate, and judge guided recalibration. By separating generation, evaluation, and revision, *HeteroForge* provides a structured approach to producing diverse educational exercises while reducing reliance on same model validation.

References

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie CK Cheung. How useful are educational questions generated by large language models? In *International Conference on Artificial Intelligence in Education*, pages 536–542. Springer, 2023.

Joy He-Yueya, Noah D Goodman, and Emma Brunskill. Evaluating and optimizing educational content with large language model judgments. *International Educational Data Mining Society*, 2024.

Euigyum Kim, Salah Khalil, and Hyo Jeong Shin. Comparing human and llm evaluations on ai-generated critical thinking items: Implications for valid applications of automatic item generation. In *Proceedings of the Second Workshop on Automated Evaluation of Learning and Assessment Content*. CEUR-WS, 2025.

Jaewook Lee, Digory Smith, Simon Woodhead, and Andrew Lan. Math multiple choice question generation via human-large language model collaboration. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 941–946, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, volume 36, pages 46534–46594, 2023.

Andrew M. Olney. Generating multiple choice questions from a textbook: Llms match human performance on most metrics. In *Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation*, pages 111–128. CEUR-WS, 2023.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 2024.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In *International conference on artificial intelligence in education*, pages 165–179. Springer, 2024.

Nicy Scaria, Silvester John Joseph Kennedy, Diksha Seth, Ananya Thakur, and Deepak N. Subramani. Harnessing structured knowledge: A concept map-based approach for high-quality multiple choice question generation with effective distractors. In *ECAI 2025 - 28th European Conference on Artificial Intelligence*, Frontiers in Artificial Intelligence and Applications, pages 4089–4096. IOS Press, 2025.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US, November 2024. Association for Computational Linguistics.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G Baraniuk. Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education*, pages 153–166. Springer, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, volume 36, pages 46595–46623, 2023.