

Can LLMs detect item flaws in the questions they have generated?

Guher Gorgun 

University of Georgia, Athens, GA, USA
gorgun@uga.edu

Abstract

As LLMs increasingly streamline item generation, ensuring the quality of their outputs remains a critical challenge. This study examines whether a multi-agent system can serve as an automated judge to detect and revise flaws in AI-generated educational assessment items, such as ambiguity, bias, or content misalignment.

1 Introduction

Generating educational content has become easier and more accessible through the use of large language models (LLMs), which can produce human-like output via chat-based interfaces with little or no programming expertise required (Kasneci et al., 2023; Brown et al., 2020). Among their educational applications, automatic item generation (AIG) has emerged as a prominent approach for developing assessment materials, with recent surveys suggesting that a substantial proportion of teachers use LLMs to support assessment development (Attard and Dingli, 2024; Hargrave et al., 2025). Although AIG can generate large numbers of assessment items rapidly, the quality of the generated items varies substantially, necessitating evaluation and revision before they can be deployed in real-world learning and assessment settings. Therefore, the primary promise of AIG (i.e., improving the efficiency of item development) is constrained by the continued need for rigorous quality evaluation and validation of generated items (Gorgun and Bulut, 2024a).

To address this challenge, we conducted a proof-of-concept study that leverages a centralized multi-agent system to generate, evaluate, and revise educational questions for a graduate-level statistics course. The system consists of an orchestrator that selects the final set of generated items and three specialized sub-agents responsible for item generation, item evaluation, and item revision. Us-

ing this framework, we investigated the extent to which LLMs can generate graduate-level statistics multiple-choice items that adhere to predefined requirements and evaluate the quality of their own generated content.

2 Related Work

To date, AIG has largely adopted a generate-then-evaluate approach, in which assessment items are first produced automatically, and their quality is assessed only after generation (e.g., Mostow et al., 2017; Gierl et al., 2024; Gorgun and Bulut, 2024a). As LLMs have made AIG more accessible to educators, the evaluation of generated items has become increasingly important for ensuring that assessment items are of sufficient quality to support valid and reliable use in learning and assessment settings (Gorgun et al., 2025). However, existing AIG evaluation practices rely heavily on human evaluators to assess the quality of each generated item individually, creating a substantial bottleneck that limits the scalability and deployment of generated items (Gorgun and Bulut, 2024a, 2025).

Furthermore, AIG research has focused exclusively on improving the grammaticality, fluency, and relevance of generated items because pre-LLM AIG systems often relied on NLP-based preprocessing techniques (e.g., syntactic parsing) and generated questions from predefined semantic or syntactic patterns. As a result, the generated items frequently exhibited rigid or template-based wording that lacked the fluency and naturalness of human-authored assessment items (Gorgun and Bulut, 2024b). LLM-based AIG has largely overcome this limitation by generating fluent, grammatically accurate, and human-like assessment questions. However, existing AIG evaluation frameworks remain limited in their ability to assess LLM-generated items because they were primarily developed for pre-LLM item generation approaches (Gorgun and

Bulut, 2024a; Kurdi et al., 2020; Flor, 2025). Therefore, LLM-generated items should be evaluated using fine-grained criteria that mirror the quality-assurance processes applied to human-authored questions, including the detection of item-writing flaws that may threaten validity and reliability (e.g., Haladyna et al., 2002; Breakall et al., 2019).

In addition to adopting more comprehensive evaluation criteria, AIG systems should consider the subject domain of the generated questions, as evaluating item quality may require domain-specific standards and expertise (Mulla and Gharpure, 2023). For example, the evaluation of mathematics word problems should account for the correctness of the underlying equations and solutions (Wang et al., 2021), whereas the evaluation of medical assessment items should consider the accuracy and appropriateness of domain-specific terminology and clinical content (Wang et al., 2007).

2.1 Current Study

These limitations in AIG highlight the need for evaluation approaches that move beyond traditional generate-then-evaluate workflows and support the automated evaluation and refinement of generated items (Gorgun and Bulut, 2024b). As LLMs increasingly produce fluent and human-like assessment questions, evaluation frameworks must be capable of identifying subtle item-writing flaws and applying domain-specific quality standards that are typically used in human review processes. Such approaches have the potential to reduce the reliance on human evaluators while improving the scalability of AIG systems.

As an initial proof-of-concept, this study explored the potential of a multi-agent system for generating, evaluating, and revising assessment items. The system consists of an item-generation agent, an item-evaluation agent, an item-revision agent, and an orchestrator agent responsible for coordinating the workflow and selecting the final set of items. Unlike conventional AIG pipelines, which treat generation and evaluation as separate stages, the proposed framework employs specialized LLM agents that assume distinct roles within the item-development process. By integrating generation, evaluation, and revision into a unified workflow, the system enables iterative improvement of assessment items prior to deployment. Specifically, the study was guided by the following research questions:

RQ1. To what extent do the evaluations produced by the multi-agent system align with human judgments of item quality?

RQ2. To what extent can a multi-agent revision process improve the quality of generated assessment items?

3 Methods

3.1 Study Design

This study investigated the feasibility of using a multi-agent system to generate, evaluate, and revise assessment items for a graduate-level statistics course. The system was designed to emulate the stages of a conventional item-development workflow by assigning distinct responsibilities to specialized LLM agents. To evaluate the consistency of the system's outputs, the complete generation pipeline was executed three times, producing a total of 60 assessment items (20 items per run). At the end of each iteration, a complete version history was maintained, including the original item generated, the number of revision cycles, the items generated in each cycle, the outcome of each quality criterion evaluated by the system for each generation cycle, the text chunk used for generating the item, and the final version of the item selected by the orchestrator for inclusion in the question bank.

3.2 Multi-Agent System Architecture

The proposed AIG systems (Figure 1) used a centralized multi-agent framework and consisted of three sub-agents and an orchestrator powered by Claude Opus 4.6, accessed through the API call. Although all agents used the same underlying model, each agent was assigned a distinct role through specialized prompting.

The item-generation agent produced questions based on predefined topics, difficulty (i.e., easy, medium, or hard), and question-type specifications (i.e., calculation, interpretation, conceptual, or application). The item-evaluation agent reviewed the generated items using a structured evaluation rubric and provided diagnostic feedback to the item-revision agent. The item-revision agent revised questions based on evaluation feedback and fed the revised questions to the item-evaluation agent for another cycle of evaluation-revision. Finally, the orchestrator agent coordinated the workflow, managed information exchange among agents, and selected the final version of each item.

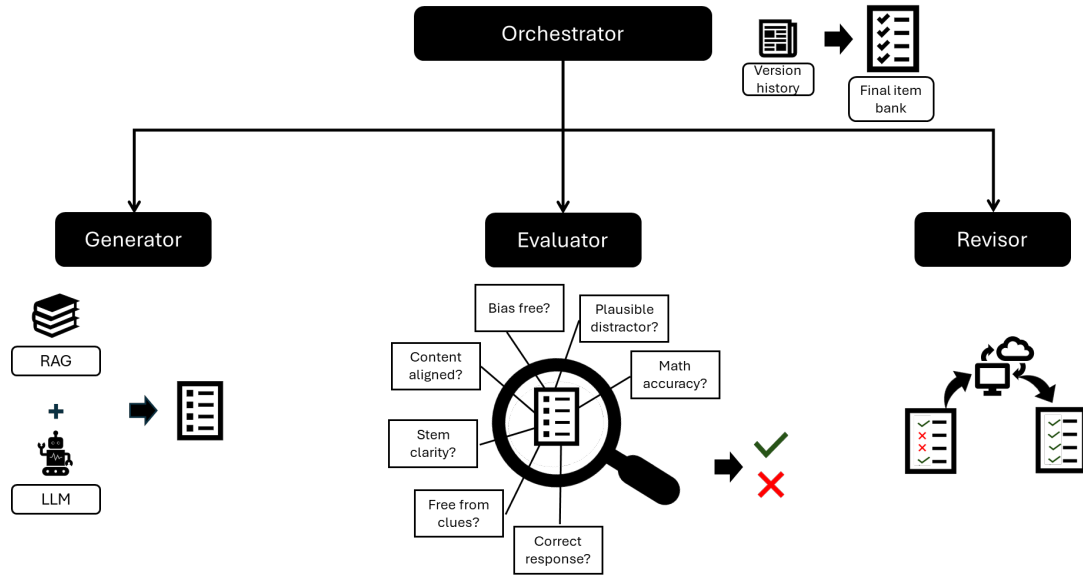


Figure 1: Workflow of the proposed multi-agent AIG framework.

3.2.1 Item-Generation Agent

We used a real-world example of a graduate-level statistics course taught at a university in the Southern United States. Using the course materials, syllabus, and learning outcomes, we developed a test blueprint covering the major topics ($N = 20$) included in the graduate-level statistics course. These topics included levels of measurement, sampling distribution, the central limit theorem, univariate and bivariate analysis, significance testing, Type I and Type II errors, effect sizes, confidence intervals, correlation, linear regression, t -test, analysis of variance (ANOVA), and chi-square analysis. We generated 20 items using the test blueprint and repeated this process three times, resulting in a total of 60 items. Item generation was guided by predefined specifications, including topic (e.g., levels of measurement), difficulty level (easy, medium, or hard), and question type (conceptual, application, interpretation, or calculation).

As previous studies highlighted the superior performance of a retrieval-augmented generation (RAG) architecture for content alignment and question quality (Gorgun and Lindner, 2026), we employed a RAG-based prompting approach during item generation. Specifically, all learning materials were indexed using a vector database, and relevant instructional content was retrieved and incorporated into the generation prompt before question generation. As our initial exploratory prompt trials revealed that the keyed response tended to be longer than the distractors, we explicitly instructed

the Item-Generator to create response options that were comparable in length and informational content. Furthermore, these initial prompt iterations revealed the need to explicitly instruct the Item-Generator to verify the accuracy of all mathematical operations. In the absence of this instruction, the Item-Evaluator sometimes struggled to evaluate the generated items, particularly when mathematical calculations contained errors. After incorporating an explicit instruction to verify mathematical operations into the generation prompt, we found that all generated items were successfully evaluated and that the Item-Evaluator correctly identified the keyed response in every case.

3.2.2 Item-Evaluation Agent

Generated items were evaluated using a 10-criterion rubric adapted from established item-writing guidelines (Haladyna et al., 2002; Breakall et al., 2019; Gierl and Lai, 2016; Moore et al., 2023). The evaluation rubric assessed 10 common item-writing flaws, including *content alignment*, *stem clarity*, *correctness of the keyed response*, *avoidance of extra information in the keyed response*, *absence of all/none of the response options*, *avoidance of absolute terms* (e.g., *all*, *always*, *every*, *only*), *absence of verbal clues*, *distractor plausibility*, *absence of bias*, and *grammatical correctness*. In addition to these domain-agnostic evaluation criteria, we included a domain-specific criterion for evaluating the accuracy of mathematical operations, as our initial prompt iterations suggested that the Item-Evaluator had difficulty assessing the quality

of questions involving mathematical calculations (e.g., calculating the mean or sum of squares from a given table). Finally, each item received an overall quality rating of *Accept*, *Minor Revision*, *Major Revision*, or *Reject*. Except for the overall quality rating, each criterion ($N = 11$) was evaluated using a binary scale indicating the presence or absence of a flaw. Criteria were coded as 0 when no flaw was identified and 1 when a flaw was present. The Item-Evaluator summed the scores for each criterion, with a total score of 0 indicating that the item satisfied all evaluation criteria. Items with a total score greater than 0 were classified as requiring revision. In addition to assigning scores, the Item-Evaluator was instructed to provide concise, actionable, and diagnostic feedback, identifying the specific flaws detected and describing what needed to be changed to address them. This feedback was then passed to the Item-Revisor and used to guide the iterative revision of the item. Thus, the evaluation process served not only to determine whether an item met the specified quality criteria but also to provide targeted guidance for correcting identified issues.

3.2.3 Item-Revision Agent

After evaluation, the revision agent received the evaluator's feedback, structured scores for each criterion, the original RAG context, and the current question text. By addressing issues mentioned in the evaluator's feedback and preserving the topic and intended question type provided to the generator, the revisor generated a revised version of the item. The orchestrator continued the evaluate-revise cycle for up to two revision rounds before selecting the final version for inclusion in the question bank.

3.3 Human Evaluation

To assess the quality of the automated evaluations, both the original and final versions of the generated items were independently reviewed by a human subject-matter expert. The reviewer was the instructor responsible for teaching the graduate-level statistics course and therefore possessed expertise in both the course content and assessment development. The expert evaluated each item using the same rubric employed by the multi-agent system. These ratings were subsequently compared with the system-generated evaluations to determine the degree of agreement between automated and human judgments of item quality.

3.4 Data Analysis

To address RQ1, we compared the evaluations produced by the multi-agent system with evaluations provided by a human expert. Agreement between the Item-Evaluator and the human expert was examined at two levels: the individual evaluation-criterion level, focusing on item-writing flaws (e.g., content alignment, distractor plausibility, or mathematical accuracy), and the overall item quality rating (i.e., accept, revise, or reject).

To answer RQ2, the effectiveness of the revision process was assessed by comparing item-quality ratings before and after revision. Descriptive statistics were used to summarize changes in rubric scores, frequencies of identified item-writing flaws, and overall item ratings across revision cycles.

4 Results

4.1 Alignment between Human Expert and Item Evaluator

Of the 60 generated items, the human expert rated 46 (77%) items as acceptable, 8 (13%) items as requiring minor revision, and 6 (10%) items as requiring major revision. Item Evaluator rated 39 (65%) items as acceptable, 14 (23%) items as requiring minor revision, and 7 (12%) items as requiring major revision. Neither the human expert nor the evaluation agent assigned a reject rating to any item.

When we compared the interrater reliability between the human expert and Item Evaluator, we found a percent agreement rate of 80% and Quadratic Weighted Kappa (QWK) value of .68, which indicated substantial agreement between the human expert and the Item-Evaluator (Landis and Koch, 1977).

We further examined the item-writing flaws identified by the human expert. The most common reason for requiring revision was the inclusion of extra information in the keyed response ($n = 10$), followed by implausible distractors ($n = 5$), the use of absolute terms (e.g., *always*, *all*, and *only*) ($n = 3$), clues in the response options ($n = 2$), and errors in the keyed response ($n = 1$).

Similar to the human expert, the Item-Evaluator identified the extra information in the keyed response as the most common reason items required revision ($n = 21$). Additional issues identified by the Item-Evaluator included the use of absolute terms ($n = 4$), implausible distractors ($n = 1$),

verbal clues ($n = 1$), mathematical inaccuracy ($n = 1$), and errors in the keyed response ($n = 1$).

4.2 Effectiveness of Item Revisor for Improving Item Quality

Of those 60 items generated, 21 (35%) items were accepted as is without requiring any revisions. During the revision process, 10 (17%) items underwent one revision cycle, whereas 29 (48%) items underwent two revision cycles. Of these 29 items going through the second revision process, 8 (13%) were ultimately accepted by the orchestrator after satisfying all evaluation criteria. However, 21 (35%) items still required revision after the second revision process.

We further analyzed the item-writing flaws identified during evaluation and examined whether the Item-Revisor was able to improve item quality by addressing these issues. Among the 18 items accepted after the first or second revision, the most frequently addressed issue was extra information in the keyed response ($n = 14$), followed by mathematical accuracy ($n = 7$), issues with the keyed response ($n = 5$), the use of absolute terms ($n = 3$), verbal clues ($n = 3$), implausible distractors ($n = 2$), and stem issues ($n = 1$).

An analysis of the items that remained in need of revision after two rounds of the evaluation–feedback–revision cycle showed that the Item-Revisor particularly struggled with items containing unnecessary information in the keyed response. Specifically, unnecessary information in the keyed response remained a persistent issue in 20 of the 21 items that required further revision. At the same time, the generation–evaluation–feedback–revision process showed that the Item-Revisor was capable of addressing other types of item-writing flaws, including mathematical inaccuracies, implausible distractors, verbal clues, absolute terms, and incorrect keyed responses. Further analysis of the items involving unnecessary information in the keyed response suggested that, in these cases, the Item-Evaluator applied a more stringent standard than the human expert, resulting in generated questions being flagged for revision more frequently or more severely than they would have been by the human expert. This suggested that the Item-Evaluator was more conservative than the human expert in evaluating the quality of the generated items.

5 Discussion

This study investigated whether a centralized multi-agent system can support the automated generation, evaluation, and revision of graduate-level statistics assessment items. The findings suggest that a structured multi-agent workflow can achieve substantial alignment with human expert judgment while enabling iterative, partially automated improvement of generated items. Furthermore, the findings highlight important design considerations for LLM-based assessment systems, particularly the need for well-specified evaluation rubrics, a combination of domain-agnostic and domain-specific criteria, and sufficiently explicit agent instructions.

With respect to RQ1, the substantial agreement between the Item-Evaluator and the human expert suggests that LLM-based evaluation can approximate expert judgment when guided by a structured rubric. Both identified overlapping item-writing flaws, with extra information in the keyed response emerging as the most frequently flagged issue for each. This convergence indicates that explicitly operationalizing item-writing principles within a rubric helps LLM-based evaluators consistently identify common sources of item-quality problems, supporting rubric-constrained prompting as a way of translating item-writing guidelines into systematically applicable criteria.

Effective automated evaluation, however, required more than generic criteria. General principles, such as avoiding verbal clues, absolute terms, implausible distractors, and unnecessary information, applied across subject areas, but other aspects of quality required domain sensitivity. Mathematical accuracy was added as a domain-specific criterion after initial trials revealed difficulty evaluating items involving statistical calculations; its inclusion let the evaluator explicitly check computational correctness rather than relying on general principles alone. This highlights that evaluation rubrics for LLM-based systems should combine domain-agnostic item-writing criteria with criteria specific to the subject area.

The comparison between the human expert and Item-Evaluator also revealed that the Item-Evaluator tended to be more conservative than the human expert, flagging unnecessary information in keyed responses substantially more often. This indicates that strong agreement at the overall rating level does not guarantee agreement on which specific flaws are present: the evaluator sometimes

relied on surface-level lexical or structural patterns, such as response length or the presence of terms like *always* or *all*, rather than contextual judgment, which in turn routed some items through unnecessary revision cycles. This stresses the importance of examining not just overall agreement but the direction and source of disagreement between automated and human judgments.

The results for RQ2 demonstrate both the promise and limits of LLM-based iterative revision. Of the 39 items that went through the revision process, 18 were accepted after one or two cycles, and the Item-Revisor successfully addressed mathematical inaccuracies, implausible distractors, verbal clues, absolute terms, and issues with keyed responses, indicating that revision is effective when a flaw is explicit and feedback is directly actionable. Extra information in keyed responses was the clear exception, persisting in 20 of the 21 items still needing revision after two cycles. Given the evaluator's conservative treatment of this criterion, this points to an interaction between evaluation and revision, a bottleneck in the feedback interface rather than a limitation of revision alone. Because the system capped revision at two cycles, future research should examine whether additional cycles improve resolution of such persistent issues.

These findings suggest that prompt and rubric design are central components of system design rather than implementation details. Structured rubrics that explicitly define each criterion, specify what constitutes a flaw, and require actionable feedback create a clearer interface between the Item-Evaluator and Item-Revisor, reducing ambiguity and making revision more effective. In general, decomposing automated assessment development into specialized agent roles allows each agent to perform a focused function and creates an iterative quality-assurance process, but the pipeline's effectiveness depends on how these components interact. That is, an overly sensitive evaluator increases the number of items routed for revision, while insufficiently specific feedback limits what the revisor can address.

Overall, these findings also imply that treating automated assessment development as a structured process (i.e., generation, evaluation, diagnostic feedback, and revision) that supports rather than replaces expert involvement. Human expertise remains essential for setting quality standards and interpreting nuanced cases, while LLM-based agents offer scalable mechanisms for generating candidate items, systematically checking their quality against

predefined criteria, and performing targeted revisions based on diagnostic feedback. Future work should investigate how human oversight can be strategically integrated into such workflows, particularly in cases where automated evaluators apply more stringent standards than human experts.

6 Conclusion

This study presented a proof-of-concept multi-agent framework for automated generation, evaluation, and revision of assessment items. LLM-based agents, organized into specialized roles and guided by a formal rubric, achieved substantial agreement with human expert judgments and supported meaningful improvements in item quality through iterative revision. Overall, centralized multi-agent architectures show promise for scalable automated assessment development, though their effectiveness depends on careful design of rubrics, instructions, and feedback mechanisms. Such systems are best understood as decision-support tools that reduce the burden of item development while preserving human oversight for nuanced quality judgments.

References

- A Attard and A Dingli. 2024. Empowering educators: Leveraging large language models to streamline content creation in education. In *ICERI2024 Proceedings*, pages 1312–1321. IATED.
- Jared Breakall, Christopher Randles, and Roy Tasker. 2019. Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice*, 20(2):369–382.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Michael Flor. 2025. *Automatic question generation*. Springer.
- Mark Gierl, Kimberly Swygert, Donna Matovinovic, Allison Kulesher, and Hollis Lai. 2024. Three sources of validation evidence needed to evaluate the quality of generated test items for medical licensure. *Teaching and Learning in Medicine*, 36(1):72–82.
- Mark J Gierl and Hollis Lai. 2016. A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice*, 35(4):6–20.
- Guher Gorgun and Okan Bulut. 2024a. Current evaluation methods are a bottleneck in automatic question

- generation. In *AI for education: Bridging innovation and responsibility at the 38th AAAI annual conference on AI*.
- Guher Gorgun and Okan Bulut. 2024b. Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Education and Information Technologies*, 29(18):24111–24142.
- Guher Gorgun and Okan Bulut. 2025. Instruction-tuned large-language models for quality control in automatic item generation: a feasibility study. *Educational Measurement: Issues and Practice*, 44(1):96–107.
- Guher Gorgun, Okan Bulut, and Bin Tan. 2025. Designing classroom assessments with generative AI. In S. Papadakis, editor, *Teaching with Artificial Intelligence: A Guide for Primary and Elementary Educators*. Taylor & Francis.
- Guher Gorgun and Marlit Lindner. 2026. Creation to evaluation: A systematic human-in-the-loop approach to automatic item generation. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME).
- Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.
- Meghan Hargrave, Douglas Fisher, and Nancy Frey. 2025. *The artificial intelligence playbook: Time-saving tools for teachers that make learning more engaging*. Corwin Press.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International journal of artificial intelligence in education*, 30(1):121–204.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.
- Steven Moore, Huy A Nguyen, Tianying Chen, and John Stamper. 2023. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In *European conference on technology enhanced learning*, pages 229–245. Springer.
- Jack Mostow, Yi-Ting Huang, Hyeju Jang, Anders Weinstein, Joe Valeri, and Donna Gates. 2017. Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children’s comprehension while reading. *Natural Language Engineering*, 23(2):245–294.
- Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.
- Weiming Wang, Tianyong Hao, and Wenyin Liu. 2007. Automatic question generation for learning evaluation in medicine. In *International conference on web-based learning*, pages 242–251. Springer.
- Zichao Wang, Andrew Lan, and Richard Baraniuk. 2021. Math word problem generation with mathematical consistency and problem context constraints. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5986–5999.