

Self-Revising Agents as Item Writers: Benchmarking Against Professional Item Writing

Steven Tang
eMetric LLC
steven@emetric.net

Zhen Li
eMetric LLC
zli@emetric.net

Richard Patz
University of California, Berkeley
rpatz@berkeley.edu

Abstract

Expert reviewers rated 156 calculus items from a self-revising multi-agent framework (Claude Sonnet 4.6 or GPT-5.4, three prompt conditions) and 26 human-written items. AI items were formative-ready nearly as often as human items (78–80% versus 88%) but summative-ready less often (26% and 10% versus 58%); a root-item reference mattered most.

1 Background

In the recent past, Automatic Item Generation (AIG) has relied on rule-based templates (Gierl and Haladyna, 2013). Contemporary Large Language Models (LLM) can now produce contextually rich items from natural-language prompts alone (Attali et al., 2022; Li et al., 2025). Li et al. (2025) found that approximately 25% of GPT-4-generated clinical items were free of major flaws and roughly half were useful starting points relative to writing from scratch.

However, when generating items from a standard alone, the model must infer content, format, difficulty, and style without guidance, and outputs often drift from the conventions of a specific course or program. Grounding generation in existing professionally written items addresses this limitation: supplying them as in-context exemplars (Brown et al., 2020) or through retrieval-grounded prompting (Lewis et al., 2020) steers the model toward established conventions and transfers design decisions that are difficult to articulate fully in a prompt.

This study investigates whether a self-revising multi-agent (SRMA) AI framework can produce useful calculus items under different prompt conditions, including conditions where a root item or a root item paired with a derived variant is provided as reference. The root items used in this study were produced by professional item writers during the 2020 pandemic whereby existing hand-written exam materials were carefully converted

into standard-aligned digital items. To support the potential of multiple forms, the item writers then created “derived” variant items from the root items. This process of creating derived items from root materials is now a seemingly natural candidate for AI to address.

2 Aims

Aim 1: Evaluate whether an SRMA pipeline can expand a bank of calculus items from human-constructed base materials at a quality comparable to professional item writers.

Aim 2: Determine how much reference material the pipeline needs, by comparing items generated from the content standard alone, from a root item, and from a root item paired with one human-written variant, and whether the second exemplar adds anything beyond the root.

Aim 3: Examine how Claude and GPT, used as backends in the pipeline, differ in output verbosity, mathematical precision, and adherence to the base item.

3 Samples

The dataset consisted of 78 items authored by professional item writers for a university-level calculus course, mapped across 6 content strands such as Functions, Limits, Derivatives, and Integrals. Items were organized into 26 item families, each anchored by a root item and two human-authored derived variants. The item writers were asked to expand the bank from the root items to support multiple test forms by creating clones or variants of each root; the guidance did not specify which. Sixteen families were multiple-choice, five were fill-in-the-blank, and five were open-ended.

4 Methods

4.1 Self-Revising Multi-Agent Pipeline

Items were generated by a self-revising multi-agent (SRMA) pipeline (Figure 1). Each item began at a

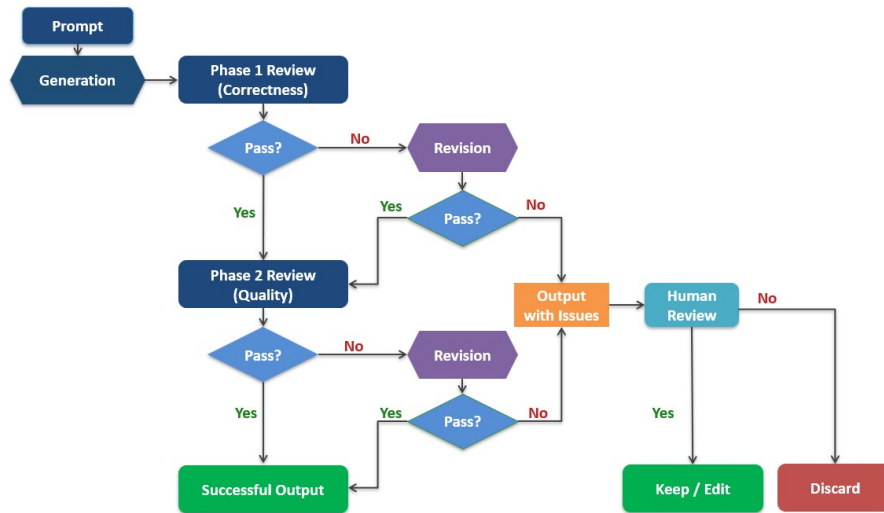


Figure 1: Diagram of SRMA Item Generation Pipeline.

Generation stage, where an LLM received a structured prompt encoding the target standard, item-writing guidelines, and reference examples from existing human-authored items. The model returned a draft item in structured format, which then entered two successive Review phases.

Phase 1 (Correctness) evaluated whether the item was mathematically sound and properly aligned to its target standard. Issues that could be resolved automatically were fixed programmatically; others triggered a targeted LLM revision followed by re-evaluation. Phase 2 (Quality) evaluated the passing items on dimensions including distractor plausibility, structural compliance, and language clarity, with the same revision-and-re-evaluate loop applied to any failures. Generated items were serialized to QTI/XML. Items failing revision in each phase were flagged for human review with a summary of remaining issues. Two generated items had a failed figure generation; these two items were discarded and replaced with a second generation attempt, which succeeded in producing the figure.

4.2 Research Design

The study employed a 2 (LLM: Claude Sonnet 4.6, GPT-5.4) $\times$ 3 (prompt condition: P1, P2, P3) fully crossed design, with human-generated items serving as a reference baseline. Two LLMs were evaluated: GPT-5.4 (OpenAI) and Claude Sonnet 4.6 (Anthropic), both demonstrating strong mathematical proficiency on established benchmarks (Wei et al., 2025; Anthropic, 2026). Three prompt conditions varied the amount of

reference information provided to each LLM, reflecting increasing levels of contextual guidance:

- P1 (Standard-Only): The LLM receives the content standard and generation instructions only, with no reference item. This condition tests the model's ability to generate a standard-aligned calculus item from scratch, without any example of what a calculus item should look like.
- P2 (Root Item): Given a root item, AI generates derived items using each LLM, enabling direct comparison between AI-generated and human-generated derived items anchored to the same root.
- P3 (Item Family): The LLM receives the content standard, a root item, and a human-generated derived item from the same family. This condition provides the fullest context, showing the model what good variation within a family looks like.

As shown in Table 1, each LLM was applied under all three prompt conditions, yielding 6 AI-generation cells of 26 items each (156 AI-generated items total). Of the 52 human-generated derived variants included, 26 (one per root item) served as the comparison baseline across conditions, and the remaining 26 (one per root item) were used as family exemplars in the P3 prompt. Generated items used the same format as their root item.

The number of items per condition was determined by the number of root items available in the human-authored calculus pool. 26 items were designated as root items, each having at least two human-authored derived items.

Prompt Condition	Source		
	Claude Sonnet 4.6	GPT-5.4	Human Expert Generated
P1	26	26	52 (26 used as comparison, 26 used as P3 exemplars)
P2	26	26	
P3	26	26	

Table 1: Conditions for item generation.

4.3 Item Evaluation

The 156 AI-generated items and the 26 human-authored comparison items (182 items in total) were rated on the 10 criteria in Appendix A: standard alignment, course appropriateness, mathematical correctness, distractor quality (multiple-choice items only), figure quality (items with a figure only), stem clarity, stem conciseness, relationship to the root item (unrelated, variant/alternate, or clone), and two holistic judgments of how much revision the item would need before use on a summative test and on a formative test (major revision, minor revision, or no revision). Reviewers also recorded open-ended comments on each item.

Every item was rated independently by two expert reviewers: the first author, and a professional item writer experienced in item review and summative test development.¹ Agreement was summarized by exact agreement, Cohen's κ (Cohen, 1960), and Gwet's AC1 (Gwet, 2008); the three measures respond differently when ratings concentrate in one category, as they do for several criteria here, so all three are reported. Every criterion on which the reviewers differed was then resolved by the first author in consultation with the second reviewer in an adjudication review. The final score on each criterion is the agreed rating or, where the reviewers differed, the adjudicated rating. A frontier LLM (Claude Fable 5.1) also rated all 182 items with the same rubric; its ratings were not used in the final scores and are reported only in the agreement analysis.

Five item-level indicators were derived from the final scores: a *fatal flaw* (a score of 0 on standard alignment, course appropriateness, or mathematical correctness); *acceptable on all criteria* (a score of 1 on alignment, appropriateness, correctness, clarity, and conciseness, and no score of 0 on distractor or figure quality where applicable; the two holistic revision judgments are not part of this indicator); *summative-ready* and

formative-ready (no revision needed for the respective use); and *clone* (interchangeable with the root item). A *composite quality* index averages the seven quality criteria, each rescaled to 0–1, excluding relationship to root and the two revision judgments, which are outcomes rather than components.

Because every family contributed one item to each condition, comparisons were paired by family: Wilcoxon signed-rank tests (Wilcoxon, 1945) on family means for pairwise comparisons between sources or conditions, Friedman tests (Friedman, 1937) across the three prompt conditions with families as blocks, and Fisher's exact tests (Fisher, 1922) for item-level rates. With 26 families the study can detect only large effects; p-values are reported descriptively without correction for multiple comparisons.

5 Results

5.1 Reviewer Agreement and Adjudication

Table 2 summarizes agreement between the two reviewers on each criterion. On the objective criteria (standard alignment, course appropriateness, and mathematical correctness) the reviewers agreed on 96–99% of items (AC1 = .95–.99). Agreement was lower on the judgment-laden criteria: 88% on conciseness, 80% on relationship to root ($\kappa = .62$) and on the formative verdict, 74% on distractor quality, and only 49% on the summative verdict ($\kappa = .15$, AC1 = .30). The summative verdict alone accounted for 93 of the 248 criterion-level disagreements, likely reflecting the two reviewers' different levels of experience in the professional authoring of summative assessments. The three agreement measures diverge where ratings concentrate in one category. On course appropriateness the reviewers disagreed on only one of the 182 items, giving 99.5% exact agreement; but because both rated nearly every item appropriate, κ , which discounts the agreement expected under that concentration, is .00, whereas AC1 applies a smaller correction and is .99. Across criteria AC1 tracks percent agreement closely without matching it. Composite quality indices computed from each reviewer's own ratings were identical for 62% of items and within .20 of each other for 96%; on average the two indices were .06 apart on the 0–1 scale, with reviewer means of .96 and .94.

¹ Janet Burnham, JLB Consulting, burnhams611@gmail.com.

Criterion	n	Agree (%)	κ	AC1	Splits
Standard alignment	182	95.6	.31	.95	8
Course appropriateness	182	99.5	.00	.99	1
Mathematical correctness	182	96.7	.39	.97	6
Distractor quality	111	73.9	.12	.70	29
Figure quality	20	80.0	-.08	.78	4
Stem clarity	182	92.9	-.03	.92	13
Stem conciseness	182	87.9	.37	.85	22
Relationship to root	182	80.2	.62	.73	36
Revision – summative	182	48.9	.15	.30	93
Revision – formative	182	80.2	.16	.78	36

Table 2: Agreement between the two expert reviewers by criterion (splits = disagreements resolved in adjudication).

Adjudication was conservative. On the evaluative criteria, 179 of the 212 disagreements (84%) were resolved to the more critical of the two ratings, including 84 of the 93 summative-verdict splits, 30 of 36 formative splits, 26 of 29 distractor splits, and all 22 conciseness splits. Relationship to root had 31 of 36 splits resolved toward being more similar to the root.

Compared with the final adjudicated scores, the AI reviewer matched 85% of all criterion-level judgments: 98–99% on standard alignment, course appropriateness, and mathematical correctness; 93% on stem clarity; 88% on conciseness; 82% on the formative verdict; 76% on relationship to root; 70% on distractor quality; and 56% on the summative verdict. Its profile thus mirrored the human reviewers': near-perfect on the objective criteria and weakest on the summative verdict.

5.2 Item Quality by Source

Table 3 reports the final scores by source. Human-authored variants had the highest composite quality (.98), followed by Claude (.94; $p = .064$) and GPT (.92; $p = .006$). Fatal flaws were rare for every source: 1 of 26 human items (3.8%), 3 of 78 Claude items (3.8%), and 2 of 78 GPT items (2.6%). The flaws differed in kind, however. The one human item with a fatal flaw was misaligned, in that it did not require the process named in the standard, whereas four of the five AI items with a fatal flaw had a mathematical flaw (a missing or duplicated key, or an ill-posed function) and one was misaligned.

	Human (n=26)	Claude (n=78)	GPT (n=78)
Criterion mean			
Standard alignment (0–1)	.96	1.00	.99
Course appropriateness (0–1)	1.00	1.00	1.00
Mathematical correctness (0–1)	1.00	.96	.99
Distractor quality (0–2) ^a	1.80	1.71	1.69
Figure quality (0–2) ^b	2.00	1.62	1.89
Stem clarity (0–1)	.96	.96	.91
Stem conciseness (0–1)	1.00	.82	.79
Relationship to root (0–2)	1.62	1.64	1.41
Revision – summative (0–2)	1.54	1.18	1.09
Revision – formative (0–2)	1.88	1.72	1.78
Composite quality (0–1)	.98	.94	.92
Indicator (% of items)			
Fatal flaw	3.8	3.8	2.6
Acceptable on all criteria	92.3	75.6	70.5
Summative-ready	57.7	25.6	10.3
Formative-ready	88.5	78.2	79.5
Clone of root	69.2	64.1	42.3

Table 3: Final scores by source. ^a Multiple-choice items only ($n = 15, 48, 48$). ^b Items with a figure only ($n = 3, 8, 9$).

The share of items acceptable on all criteria was 92% for human items, 76% for Claude (Fisher's exact $p = .09$), and 71% for GPT ($p = .03$). The criterion means in Table 3 locate the differences: AI items were less concise (.82 and .79 versus 1.00), had slightly weaker distractor sets (1.71 and 1.69 versus 1.80), and GPT stems were less often clear (.91 versus .96), whereas alignment was near ceiling for every source (.96–1.00). Conciseness accounted for most of the shortfall: of the 42 AI items that did not meet every criterion, 26 missed on conciseness alone and 4 more on conciseness together with an unclear stem; the remaining 12 involved an unclear stem (5), a mathematical flaw (4), defective distractors (2), or misalignment (1). Claude's items were as close to their roots as the human variants (64% versus 69% clones), while GPT's were farther (42%). These comparisons pool the three prompt conditions; Section 5.3 compares the root-conditioned prompts alone with the human variants.

The sources diverged sharply on readiness. 58% of human variants were judged usable on a summative test without revision, against 26% of Claude items ($p = .008$) and 10% of GPT items ($p < .001$). Major revision was required for 6 Claude items, 1 GPT item, and 1 human item. For formative use the gap largely disappeared: 78% (Claude), 80% (GPT), and 88% (human) of items

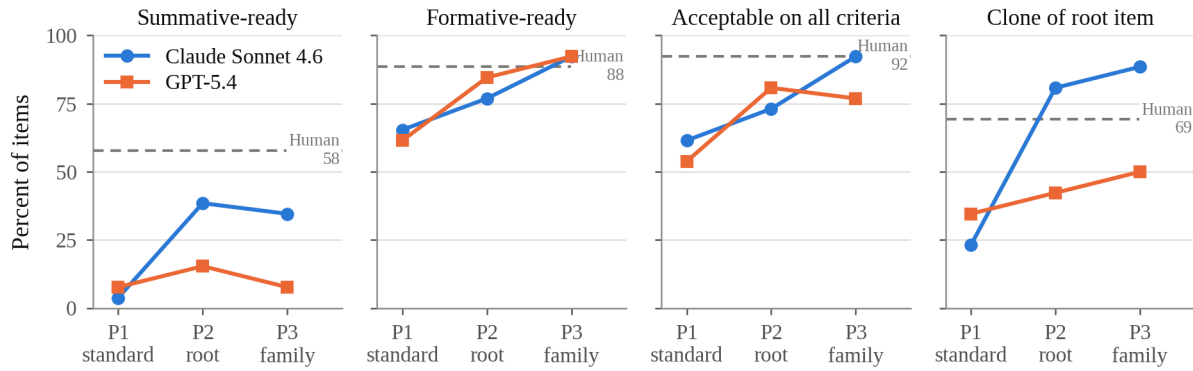


Figure 2: Summative readiness, formative readiness, share of items acceptable on all criteria, and clone rate by prompt condition and model (n = 26 items per point; dashed lines mark the human-written variants).

needed no revision, and none of the pairwise differences was statistically significant (all $p \geq .24$).

5.3 Effect of Reference Material

Of the two design factors, the amount of reference material had the larger effect, moving every indicator more than the choice of model did (Table 4, Figure 2). Pooled over the two models, moving from the standard alone (P1) to a root item (P2) to an item family (P3) raised the share of items acceptable on all criteria from 58% to 77% to 85%, formative readiness from 63% to 81% to 92% (Friedman $p < .001$), the clone rate from 29% to 62% to 69% ($p < .001$), and mean conciseness from .67 to .85 to .90 ($p = .003$).

Condition	Composite	Fatal	All accept.	Summ. ready	Form. ready	Clone	Concise
Claude P1	.91	0.0	61.5	3.8	65.4	23.1	.65
Claude P2	.94	7.7	73.1	38.5	76.9	80.8	.85
Claude P3	.97	3.8	92.3	34.6	92.3	88.5	.96
GPT P1	.89	3.8	53.8	7.7	61.5	34.6	.69
GPT P2	.94	0.0	80.8	15.4	84.6	42.3	.85
GPT P3	.94	3.8	76.9	7.7	92.3	50.0	.85
Human	.98	3.8	92.3	57.7	88.5	69.2	1.00

Table 4: Indicators by generation condition (n = 26 items per row; All accept. = acceptable on all criteria).

For Claude, a root item transformed the output: summative readiness rose from 4% in P1 to 38% in P2 and 35% in P3 ($p = .002$), the clone rate from 23% to 81% to 88% ($p < .001$), the share acceptable on all criteria from 62% to 73% to 92%, and formative readiness from 65% to 77% to 92% ($p = .025$). For GPT, reference material improved the share acceptable on all criteria (54%, 81%, 77%), composite quality (.89, .94, .94; $p = .031$), and formative readiness (62%, 85%, 92%; $p = .009$) but not summative readiness (8%, 15%, 8%; $p = .37$),

and its items remained farther from the root (35%, 42%, 50% clones; $p = .22$).

Two further patterns matter for practice. First, the second exemplar improved criterion-level quality rather than readiness: from P2 to P3, the share of Claude's items acceptable on all criteria rose from 73% to 92% and its formative readiness from 77% to 92%, but its summative readiness did not (38% versus 35%). Second, reference material did nothing for mathematical correctness: fatal flaws occurred at a low rate in every condition (pooled 1.9%, 3.8%, 3.8%; Claude 0%, 7.7%, 3.8%; GPT 3.8%, 0%, 3.8%), and all three of Claude's mathematical errors occurred in items generated with a reference item. In the strongest AI condition, Claude with an item family, 92% of items were acceptable on all criteria, 88% were clones of their root, and 92% were ready for formative use, yet only 35% were ready for summative use, compared with 58% of the human variants.

The human variants also provide a like-for-like comparison for the root-conditioned prompts. The item writers and the P2 and P3 runs started from the same root items; the writers were asked for clones or variants of each root, the prompts asked for an item variant, and neither specified how close to the root the new item should stay. The comparison therefore shows where professionals and models each landed on the clone-variant spectrum given the same starting material. With the root supplied, Claude's items were classified as clones at least as often as the human variants (81% in P2 and 88% in P3 versus 69%; $p = .37$ and $.10$), while GPT's items departed from the root more often (42% and 50% clones; $p = .07$ and $.13$). On the readiness verdicts, Claude's items were summative-ready less often

than the human variants (38% and 35% versus 58%), a difference that was not statistically significant ($p = .13$ and $.11$), whereas GPT's items were summative-ready much less often (15% and 8%; $p = .002$ and $< .001$); formative readiness was at the human level for both models in P3 (92% versus 88%).

5.4 Model Differences

Compared directly with each other (Aim 3), the two models did not differ in composite quality, .94 for Claude and .92 for GPT ($p = .38$), but their profiles differed in several respects. Claude's items were summative-ready more often than GPT's, 26% against 10% over all conditions ($p = .017$), and the difference held within the two root-conditioned prompts, 38% against 15% in P2 ($p = .034$) and 35% against 8% in P3 ($p = .020$). Claude's items were also closer to the root: 64% were clones, against 42% of GPT's ($p = .011$). GPT's items, in turn, contained fewer mathematical errors, 1 against Claude's 3, and fewer required major revision, 1 against 6 ($p = .12$). At the same time, GPT's stems were more often unclear, 7 against 3, and drew more verbosity comments from the reviewers, 42 against 27 mentions per 100 items summed over the three raters. On formative readiness the models were indistinguishable, at 78% and 80%.

5.5 Item Format

The differences reported above depended on item format (Table 5). For multiple-choice items, both models approached the human benchmark on formative readiness (Claude 85%, GPT 94%, human 93%) and on the share acceptable on all criteria (85%, 90%, 100%). For open-ended items, AI performance fell away: formative readiness dropped to 53% (Claude) and 47% (GPT) against 83% for human items, the share acceptable on all criteria to 47% and 33%, and mean conciseness to .47 and .40; two of Claude's major-revision items were open-ended items generated from the standard alone. Fill-in-the-blank items separated the models, with Claude reaching 80% formative and 47% summative readiness against GPT's 67% and 13%. These cells are small (15 AI items and 5–6 human items per format) and were not tested.

Format / source	n	All accept.	Summ.-ready	Form.-ready
Multiple choice				
Human	15	100.0	60.0	93.3
Claude	48	85.4	22.9	85.4
GPT	48	89.6	12.5	93.8
Fill-in-the-blank				
Human	5	80.0	80.0	80.0
Claude	15	73.3	46.7	80.0
GPT	15	46.7	13.3	66.7
Open-ended				
Human	6	83.3	33.3	83.3
Claude	15	46.7	13.3	53.3
GPT	15	33.3	0.0	46.7

Table 5: Item-level indicators by item format and source (percent of items; All accept. = acceptable on all criteria).

6 Discussion

With respect to Aim 1, the results indicate that the pipeline can contribute usable items to a calculus bank. AI items were as free of fatal flaws as the professional variants (3–4% for all sources), were acceptable on all criteria about three-quarters of the time, and were judged formative-ready at rates close to those of the human items (78–80% versus 88%). The gap in summative readiness consisted mainly of minor revisions, such as trimming stems and strengthening distractor sets, rather than major ones. In the conditions most comparable to the item writers' own task, the root-conditioned prompts, Claude's items were summative-ready 35–38% of the time against 58% for the human variants, a difference that was not statistically significant, whereas GPT's items remained well below (8–15%). The few mathematical errors that occurred had passed the pipeline's own review phases, an indication that AI review of AI-generated content, while useful, is not a complete substitute for human checking.

Among the factors examined, the amount of reference material had the largest effect (Aim 2). Items generated from the standard alone tended to restate definitions and scaffold the solution method, whereas items generated from a root item followed the conventions of the bank more closely. Adding a second exemplar was associated with a higher share of items acceptable on all criteria but not with higher summative readiness. The root items were written by professionals and served as good examples, but they had not been screened to be flawless, and the human variants derived from

them were judged summative-ready 58% of the time. Because generation conditions on the root, a root that would itself need revision may carry its shortcomings into the generated item, so the summative-readiness rates observed here may be lower than those attainable with roots rigorously curated for high-stakes summative purposes.

The two models differed in profile more than in overall quality (Aim 3). Claude’s items were more often close to the root and more often summative-ready; GPT’s items contained fewer mathematical errors but were more often verbose and less closely tied to the root. Which profile is preferable depends on whether the goal is parallel forms or broader coverage. A pattern common to both models was the inclusion of material that a summative item would omit, such as hints, restated definitions, and decorative prose. This suggests that the models’ default notion of an item is closer to a formative than a summative one. Whether that default is a limitation depends on the intended use, and how prompting strategies can best be tuned to different uses remains open for future work.

References

- Anthropic. 2026. [Claude Sonnet 4.6 system card](#).
- Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A. von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Ronald A. Fisher. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94.
- Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.
- Mark J. Gierl and Thomas M. Haladyna, editors. 2013. *Automatic Item Generation: Theory and Practice*. Routledge, New York, NY.
- Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Jiaxuan Li, Saed Rezayi, Peter Baldwin, Polina Harik, and Victoria Yaneva. 2025. [Towards reliable generation of clinical chart items: A counterfactual reasoning approach with large language models](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 142–153, Pittsburgh, Pennsylvania, USA. National Council on Measurement in Education (NCME).
- Hu Wei, Ze Xu, Boyu Yang, Linlin Miao, Weiqi Zhai, Yihan Li, Zixuan Li, Zhijun Wang, Boya Wang, Jianwei Yu, Jialing Yuan, Xiaoyue Zhang, Cheng He, Minglei Chen, Zifan Zhang, Qianhui Li, Wei Wang, and Xiang Xu. 2025. [SKYLENAGE technical report: Mathematical reasoning and contest-innovation benchmarks for multi-level math evaluation](#). Preprint, arXiv:2510.01241.
- Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

A Item Evaluation Criteria

Table A.1 lists the criteria applied by the expert reviewers to every AI-generated and human-authored comparison item (see Section 4.3).

Criterion	What Reviewers Evaluate	Scale
Standard Alignment	Item requires the cognitive process specified in the standard or learning objective	0 = Not aligned 1 = Aligned
Course Appropriateness	Content and demand suitable for university calculus students	0 = Inappropriate 1 = Appropriate
Mathematical Correctness	The mathematical content, terms, definitions, and symbols used in the item must be entirely correct. For Multiple Choice Item: – Key is unambiguously correct; – All distractors are definitively incorrect	0 = Error present 1 = Correct
Distractor Quality (Set) (Multiple Choice items only)	Acceptable distractors are unambiguously incorrect. Good distractors are acceptable but also plausible under an identifiable misconception The set of distractors is: Defective if any element is not Acceptable Acceptable if each element is only Acceptable Good if each element is Acceptable, at least one is Good, and no option (including the correct one) stands out suggestively (i.e., is an ‘outlier’) Excellent if each element is Good and there are no outliers	0 = Defective 1 = Acceptable 2 = Good or Excellent
Figure Quality	Figure (if any) is mathematically correct, meaningful and displayed well.	0 = Defective 1 = Usable with minor revision 2 = Good
Stem Clarity	Stem language is clear for the math task – Question is unambiguous	0 = Unclear 1 = Appropriate
Stem Conciseness	Stem language is concise for the math task – No decorative or redundant prose	0 = Verbose 1 = Appropriate
Relationship to “Root” Item	Clone that is interchangeable with root (same standard, same concept, same skills, same difficulty level); Variant/Alternate that addresses the same standard but uses a similar or distinctly different conceptual approach, required skills, or difficulty; Unrelated	0 = Unrelated 1 = Variant/Alternate 2 = Clone
Open Ended	Make a few observations about this item, anything that stands out as particularly strong or in need of revision.	
Revision for Summative Use	How much revision is needed before this item could be used on a summative test?	0 = Needs major revision 1 = Needs minor revision 2 = No revision needed
Revision for Formative Use	How much revision is needed before this item could be used on a formative test?	0 = Needs major revision 1 = Needs minor revision 2 = No revision needed

Table A.1: Item evaluation criteria applied by expert human reviewers.