Improving Multiple Choice Questions Using Synthetic Data

Maya Bialik [1,2] and Will Cummings [2]

[1] Boston University Wheelock College of Education and Human Development
mbialik@bu.edu
ORCID: 0009-0002-9860-0508
[2] QuestionWell
will@questionwell.ai

Abstract. Large language models (LLMs) often generate multiple-choice questions with item-writing flaws even when given explicit rubrics and examples. We present a rubric-driven post-training pipeline for automatically verifiable MCQ flaws and evaluate the resulting QuestionWell model against `gpt-4.1-nano` and `gemini-2.5-flash`. The final rubric flags five flaws: longest-answer cueing, absolute terms such as "always" or "never," fill-in-the-blank format, true/false degeneration, and "all/none of the above" options. Starting from `gemini-2.5-flash`, we generated 1,680 ranked pairs with $n = 2$ sampling, used the accepted generations for a 15-epoch supervised warmup, and then ran 10 epochs of direct preference optimization with $\beta = 0.1$ on the same pairs. On 50 English quiz generations per model, QuestionWell averaged 96.0% unflawed questions per quiz, versus 69.2% for `gemini-2.5-flash` and 55.1% for `gpt-4.1-nano`. It also reached at least 90% flaw-free performance in 48 of 50 quizzes, compared with 5 of 50 for Gemini and 1 of 50 for GPT. These results show that targeted rubric-driven training can substantially improve both average MCQ quality and generation consistency.

Keywords: multiple-choice questions · synthetic data · item-writing flaws · large language models · post-training

1 Introduction

Multiple-choice questions remain a cornerstone of educational assessment, yet their quality depends critically on established item-writing principles. Decades of psychometric research have identified specific flaws, such as "all/none of the above" options, fill-in-the-blank formats, and length-based answer cues, that systematically compromise item validity by introducing construct-irrelevant variance and providing unintended clues to test-takers [2]. These flaws allow test-wise students to identify correct answers without mastering the underlying content, undermining both the formative and summative value of the assessment [1]. For educators, this is not merely a measurement concern. Teachers increasingly rely on auto-generated quizzes for low-stakes formative practice—retrieval exercises, comprehension checks, homework sets—where question volume matters

and manual review of every item is impractical [6]. When the generator itself is unreliable, the burden of quality control falls back on the teacher, eroding the time savings that motivated automation in the first place. Despite this well-documented knowledge base, modern large language models consistently produce questions exhibiting these same problematic patterns.

The persistence of item-writing flaws in LLM-generated questions presents a puzzle: these errors are easy to detect automatically, and models can understand explicit instructions about what to avoid. Yet detailed rubrics, examples, and prohibitions still fail to eliminate them. In our preliminary experiments, models continued to generate longest-answer cues and "all of the above" options even when shown 10–20 high-quality exemplars.

This phenomenon appears across models and providers. The "longest answer is correct" flaw, in particular, recurs with striking consistency across systems of similar capability, suggesting that the problem reflects pre-training data biases rather than a simple failure of instruction-following. Although frontier reasoning models can avoid some of these issues more reliably, they remain too expensive and high-latency for many production workflows.

We hypothesize that these flaws are embedded in pre-training data. Rather than manually curating thousands of high-quality examples, we leverage the automatic verifiability of these flaws to construct scalable preference signals, training the model to become a more reliable partner for educators generating assessments at scale.

2 Related Work

Research on MCQ quality spans psychometrics, automated item analysis, and LLM-based question generation. Prior work catalogues recurring item-writing flaws and shows that they distort difficulty and discrimination, motivating automated checks for option-set pathologies and surface cues [2,3]. Automatic question generation has a long history in educational AI; Kurdi et al. provide a comprehensive review of methods through 2019, noting persistent challenges in controlling item quality [4]. Moore et al. developed a rule-based method to automatically detect 19 common item-writing flaws in student-generated MCQs, outperforming GPT-4 on the same task [5]. Recent work has also examined LLM-generated MCQs specifically: Dijkstra et al. fine-tuned GPT-3 on text–quiz pairs and found that distractor quality remained the primary bottleneck [6], while Tan et al. review automatic item generation techniques leveraging LLMs and find that generated items still require substantial human revision before classroom use [7]. Our work extends that evaluation paradigm into post-training: instead of only filtering outputs after generation, we use automatically scored synthetic data to move the model away from flaw-inducing modes.

3 Methodology

We used a prompt that asks the model to generate multiple-choice questions from a source text. We scored each generated quiz by the percentage of questions that were free of detectable item-writing flaws. Following prior work on automated MCQ evaluation [5], we restricted the rubric to automatically checkable heuristics so that large numbers of generations could be audited consistently.

3.1 Rubric

For a quiz with C generated questions, let F be the number of questions with at least one detected flaw. We define quiz quality as

$$\text{Score} = \frac{C - F}{C},$$

so a score of 1 indicates a fully flaw-free quiz.

The final evaluation marks a question as flawed if any of the following conditions hold:

- the correct answer is the longest choice by a large enough margin to create a cueing effect;
- the stem or choices contain absolute terms such as "always" or "never";
- the item is written as fill-in-the-blank rather than standard multiple choice;
- the item degenerates into true/false format; or
- one of the options is "all of the above" or "none of the above."

This rubric matches the final evaluation and is intentionally restricted to flaws that can be checked reliably at scale.

3.2 Training Procedure

We first generated candidate quizzes and scored them with the automated rubric above. Specifically, we sampled $n = 2$ generations per prompt, ranked the resulting 1,680 pairs with our automated grader, and retained the accepted/rejected structure for training. We then trained QuestionWell to reduce the measured flaws while monitoring separate guardrails for source similarity, Lexile level, question count, learning-outcome alignment, and Bloom's taxonomy distribution.

The final production model used a two-stage post-training recipe:

- **Supervised warmup.** We ran supervised tuning on top of `gemini-2.5-flash` for 15 epochs using the accepted side of those same 1,680 ranked pairs. Intermediate checkpoints were enabled, the selected default checkpoint was 10, the learning-rate multiplier was 1, and the adapter size was 1.
- **DPO stage.** We then ran a DPO stage for 10 epochs on the same ranked pairs used to construct the warmup data. Intermediate checkpoints were enabled, the selected default checkpoint was 13, the learning-rate multiplier remained 1, the adapter size remained 1, and the DPO β parameter was 0.1.

Table 1. Final evaluation summary.

Treatment	Mean flaw-free rate	Quizzes at ≥90%
`gpt-4.1-nano`	55.1%	1 / 50
`gemini-2.5-flash`	69.2%	5 / 50
QuestionWell	96.0%	48 / 50

The final model reported in Section 4 is the production version of this two-stage flaw-reduction system.

3.3 Evaluation

We performed the final evaluation with an improved longest-answer heuristic that only fires when the correct option is substantially longer than the longest distractor. Let

$$L_c^{(j)} := |O_{c^{(j)}}^{(j)}|, \qquad L_d^{(j)} := \max_{i \neq c^{(j)}} |O_i^{(j)}|$$

$$\text{LongestCorrect}(j) = (L_d > 0) \wedge \left(\frac{L_c}{L_d} \geq 1.2 \right) \wedge (L_c - L_d \geq 5)$$

We then generated 50 English-language quizzes per model from the same prompt and scored each quiz using the final rubric. The comparison includes three systems: `gpt-4.1-nano`, `gemini-2.5-flash`, and QuestionWell. We restricted the evaluation to English outputs so that manual inspection would be directly comparable across models.

- Quiz quality: the fraction of questions in each quiz with no detected flaw.
- Consistency: how many generated quizzes stay above increasingly strict flaw-free thresholds.
- Guardrail attributes monitored during product evaluation: text similarity to the source passage, Lexile level, number of questions, learning-outcome alignment, and Bloom's taxonomy distribution.

4 Results

The final comparison emphasizes two properties: average quiz quality and consistency across repeated generations. Figure 1 summarizes both. Across 50 sampled English quizzes per model, QuestionWell achieved 96.0% unflawed questions per quiz on average, compared with 69.2% for `gemini-2.5-flash` and 55.1% for `gpt-4.1-nano`. The cumulative-threshold curve shows the more important practical result: QuestionWell remains near-perfect through stringent thresholds, while both comparison models deteriorate much earlier.

By inspection of Figure 1, 48 of 50 QuestionWell quizzes are at least 90% flaw-free, compared with 5 of 50 Gemini 2.5 Flash quizzes and only 1 of 50 GPT-4.1 nano quizzes. This gap matters more than the mean scores alone because quiz-generation systems are experienced one sample at a time.

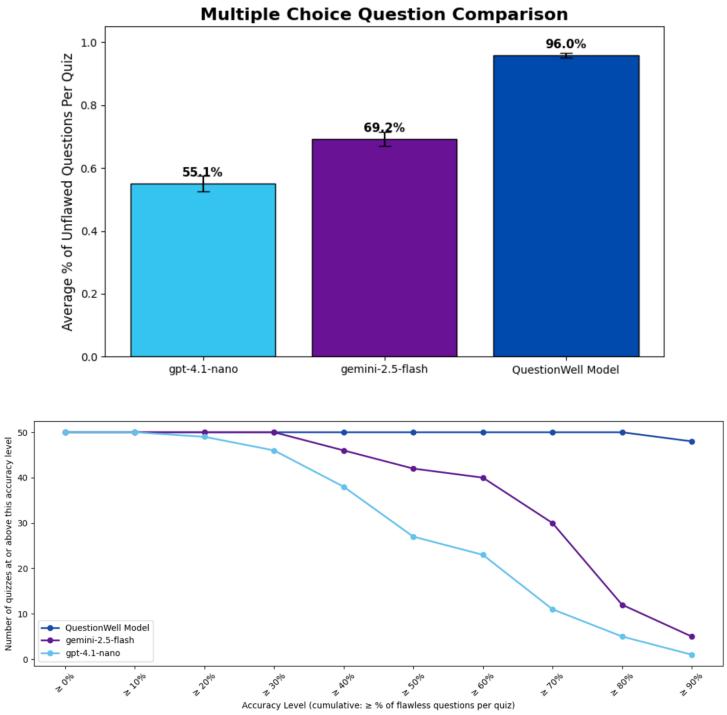

Fig. 1. Final comparison results. Top: average quiz quality. Bottom: cumulative consistency, defined as the number of generated quizzes whose flaw-free rate meets or exceeds each threshold.

The evaluation additionally tracked non-target attributes such as similarity to the source passage, reading-level adherence, number of questions, learning-outcome alignment, and Bloom's taxonomy distribution. Table 2 shows that QuestionWell preserved question count relative to `gemini-2.5-flash` (11.28 versus 11.24 questions per quiz on average), reached perfect learning-outcome alignment on this sample, and shifted Bloom shares modestly toward Apply.

5 Limitations

This study has several limitations. The comparison uses only 50 outputs per model from one prompt configuration, so the results are descriptive rather than definitive. The evaluation is also restricted to English-language generations, and all reported quality and guardrail metrics are automated proxies rather than exhaustive measures of question quality. Finally, one `gpt-4.1-nano` run is missing a context-similarity value, so that summary is based on 49 rather than 50 outputs.

Table 2. Guardrail metrics and Bloom-category shares by model. Guardrail entries report mean with population standard deviation in parentheses. Context similarity for `gpt-4.1-nano` is based on 49 non-missing runs. Bloom entries are percentages and may not sum to exactly 100 because of rounding.

Treatment	Context sim.	Reading score	LO align.	Questions
`gpt-4.1-nano`	0.565 (0.098)	0.704 (0.291)	0.576 (0.401)	9.42 (3.99)
`gemini-2.5-flash`	0.609 (0.054)	0.733 (0.288)	0.981 (0.056)	11.24 (5.06)
QuestionWell	0.554 (0.051)	0.649 (0.275)	1.000 (0.000)	11.28 (5.22)

Treatment	Rem.	Und.	App.	Ana.	Eval.	Cre.
`gpt-4.1-nano`	40.4	33.8	6.6	12.8	4.7	1.7
`gemini-2.5-flash`	24.5	36.9	16.7	13.5	5.8	2.7
QuestionWell	19.1	40.2	19.4	13.5	5.2	2.6

6 Conclusion

The final evaluation supports the paper's central claim: targeted rubric-driven training can materially reduce common MCQ item-writing flaws. On 50 English quiz generations from a shared prompt, QuestionWell substantially outperformed both comparison models in average flaw-free rate and in consistency above the 90% threshold. Future work should expand the evaluation to additional languages and broader item-writing criteria. By moving flaw detection from a post-hoc filter to a training signal, this approach brings auto-generated assessments closer to the reliability teachers need before they can trust them at scale.

References

1. Downing, S.M.: The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Advances in Health Sciences Education **10**(2), 133–143 (2005)
2. Haladyna, T.M., Downing, S.M.: A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education **2**(1), 37–50 (1989)
3. Schmucker, R., Moore, S.: The impact of item-writing flaws on difficulty and discrimination in item response theory. arXiv preprint arXiv:2503.10533 (2025). https://arxiv.org/abs/2503.10533
4. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education **30**(1), 121–204 (2020)
5. Moore, S., Nguyen, H., Chen, T., Stamper, J.: Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods. In: European Conference on Technology Enhanced Learning (2023). https://arxiv.org/abs/2307.08161
6. R. Dijkstra, Z. Genç, S. Kayal, J. Kamps, et al., "Reading Comprehension Quiz Generation using Generative Pre-trained Transformers," in *iTextbooks@AIED*, pp. 4–17, 2022.

7. Tan, B., Armoush, N., Mazzullo, E., Bulut, O., Gierl, M.J.: A review of automatic item generation techniques leveraging large language models. International Journal of Assessment Tools in Education **12**(2), 317–340 (2025)