Evaluating Large Language Models of Rasch Calibrated Multiple-Choice Questions: An Error-Bound Pareto Frontier Framework for Model Selection

Hariz Liew Zhen Wei[1,a,*] and Che Yee Lye[1,b]

[1] Teaching & Learning Centre, Singapore University of Social Sciences, 463 Clementi Road, Singapore 599494

[a,*] harizliewzw@suss.edu.sg, [b] cylye@suss.edu.sg

Abstract. The potential of Large Language Models (LLMs) to support scalable multiple-choice question (MCQ) generation is promising. However, institutes of higher learning (IHLs) struggle with decision-making due to rapid model churn. Static benchmarks also become obsolete quickly, leaving IHLs without a durable strategy. Additionally, a key but underemphasised challenge is that LLM accuracy is evaluated on item sets with uncalibrated difficulty, resulting in misleading accuracy comparisons across models. This paper introduces a version-robust Error-Bound Pareto Frontier Framework for selecting LLMs for university-level MCQs, grounded in psychometrically calibrated assessment data. Using a dataset of 1,020 MCQs across Mathematics, Statistics and Psychology, item difficulty was calibrated using the Rasch model, with item fit analysis applied to ensure measurement validity before LLM evaluation. Difficulty annotations derived from Rasch-calibrated item parameters were then used to demonstrate the framework. By establishing strict error tolerance thresholds (Guardrails) as a priority before optimising for cost and latency, results indicate that while cutting-edge models such as o3-mini and gpt5-nano dominated accuracy benchmarks, cost-efficient alternatives such as DeepSeek-V3 defined part of the Pareto frontier in selected settings. Grounded in psychometric calibration, the framework offers IHLs a practical method for model selection, maintaining strict response accuracy amid rapid model evolution.

Keywords: large language models, multiple-choice questions, Rasch measurement model, model evaluation, Pareto Frontier

1 Introduction

Multiple-choice questions (MCQs) remain central in institutes of higher learning (IHLs) because they are objective, efficient, and scalable (Moss, 2001; Gierl et al., 2017; Shin et al., 2019; Oc & Hassen, 2025). They are also well-suited for integration into adaptive learning systems, which rely heavily on a large pool of validated items

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_22

that can be auto marked, to deliver effective personalisation in learning (Imhof et al., 2020; Khosravi et al., 2020). Building and maintaining such item banks, however, is costly and labor-intensive, requiring sustained subject-matter expertise and alignment with learning outcome (Haladyna et al., 2002; Gierl et al., 2017, Mucciaccia et al., 2025). Recent studies have shown evidence that Large Language Models (LLMs) can help scale MCQ generation workflows, if they can be deployed reliably and cost-effectively (Mucciaccia et al., 2025; Nikolovski et al., 2025).

While LLMs have been proposed as a way to scale MCQ generation workflows, evidence remains mixed and fragmented across question format (Li et al., 2024), subject areas (Savelka et al., 2023; Ramesh et al., 2024), and cognitive complexity (Hwang et al., 2024). LLMs are often used to answer long-form generation or open-ended questions (Li et al., 2024), and their consistency in handling MCQs across different subject areas and cognitive levels, remains an open question. Compounding this issue, rapid LLM turnover makes static benchmarks short-lived and difficult to operationalise in real deployments. What is missing is not another leaderboard but guidance that helps IHLs make decisions that remain valid as models evolve.

A key but underemphasised challenge is that LLM accuracy is evaluated on item sets in which their difficulty may not be psychometrically calibrated, limiting meaningful comparison and interpretation (Wang et al., 2023). The Rasch measurement model addresses this by estimating item difficulty on a common scale, enabling direct comparison across LLM performance across models and MCQ sets.

To address these gaps, this study employed the Rasch measurement model to calibrate MCQs and introduced the Error-Bound Pareto Frontier, an evaluation framework based on Pareto optimality (Pareto, 1906/1971)**, to** support reliable deployment of MCQ generation in IHLs, even as underlying models rapidly evolve. The framework was grounded in a real-world context, drawing on a large item bank of 1,020 Rasch-calibrated MCQs from university-level Mathematics, Psychology and Statistics, accompanied by item difficulty and learning-progression annotations. This study contributes to a decision process that prioritised non-negotiable operational constraints, specifically error tolerances (the "Error Bound") before evaluating trade-offs in latency and cost (the "Pareto Frontier"). This selection method could be integrated with practical deployment patterns to ensure long-term comparability and quality control.

2 Psychometric Principles and Large Language Model Evaluation

LLM evaluation for assessment tasks such as MCQs often mirrors Classical Test Theory (CTT) by aggregating performance across a fixed set of test items to produce a single accuracy score (Ye et al., 2025). Similar to CTT, this approach treats all items as equally informative and does not explicitly model item difficulty or measurement error.

However, closer examination reveals that prevailing LLM evaluation benchmarks diverge substantially from modern test theories such as the Rasch measurement model (Wang et al., 2023). The following section outlines the need for a psychometrically rigorous approach to LLM evaluation.

2.1 Large Language Model Evaluation

LLMs are advanced machine learning systems trained to predict and generate text based on large-scale corpora and are built primarily based on the transformer architectures (Vaswani et al., 2017). Their success across a range of tasks has led to widespread adoption in natural language tasks, including applications in education such as question generation. LLM performance is typically evaluated based on a set of tasks, such as MCQs, reading comprehension or reasoning. These benchmarks produce leaderboard-style results that are easy to compare by using aggregate performance such as overall accuracy. However, such evaluations are often based on static datasets that may not reflect the needs of specific domains (Eriksson et al., 2025). Additionally, model accuracy in question answering tasks are usually reported without considering the difficulty or psychometric validity of test items (Rodriguez et al., 2021). Recent work has expanded evaluation beyond pure task success to include system-level constraints such as inference cost and latency as part of the evaluation and decision making. As a result, models are increasingly compared along Pareto frontiers, where optimal models depend on the trade-offs between accuracy, cost, and latency (Erdil, 2025). There is also growing interest in using LLMs for automated item generation, particularly for MCQs. Studies have shown promising results in generating questions, keys and distractors, but evaluation often lacks psychometric principles, limiting reliability in format assessment contexts (Bitew et al., 2023; Kumar et al., 2023; Mucciaccia et al., 2025; Raina & Gales, 2022).

2.2 Rasch Measurement Model

Central to psychometric principles is the prioritisation of item quality through rigorous item analysis. Advanced statistical models, such as Rasch measurement model, are widely adopted to estimate latent traits, evaluate item performance and assess model fit (Aryadoust, et al., 2020). When applied to LLM evaluation, Rasch analysis moves beyond leaderboard-style benchmarking practices that rely on aggregate accuracy scores computed over fixed and uncalibrated items (Ye et al., 2025). It examines item informativeness and ensures that performance comparisons are grounded in calibrated item difficulty. This approach differs fundamentally from most LLM benchmarks, which typically assume item equivalence, and provide limited insight into how performance varies with task difficulty, leading to scores that are non-comparable across benchmarks (Wang et al., 2023) and poorly aligned with deployment decision-making. In contrast, Rasch measurement explicitly models latent construct ability and item difficulty on a common scale, enabling invariant comparison across models, test forms and evaluation contexts.

3 Methods

3.1 Datasets

The datasets for this study comprised 1,020 MCQs derived from one Mathematics (i.e., Course 1), two Psychology (i.e., combined into Course 2) and one Statistic (i.e., Course 3) courses. The data points extracted from the database include question text, key or correct option, distractors or incorrect options, difficulty estimate and difficulty level.

To operationalise item difficulty for subsequent LLM evaluation, the MCQs for each course were calibrated using the Rasch measurement model (Wright & Masters, 1982) to examine their item fit and determine difficulty estimates (in logits). These difficulty estimates were then recategorised into three difficulty levels (i.e., Levels 1–3) within each course to support stratified reporting and analysis. Breakpoints between levels were determined using the Jenks natural breaks method (Jenks, 1967), which selects cut points that minimise within-level dispersion while maximising separation across levels, producing course-specific strata that adapt to non-uniform or clustered difficulty distributions. Because the cut points are estimated separately for each course, these difficulty levels should be interpreted as within-course difficulty bands and are used to inform difficulty-aware model evaluation and deployment decisions.

Although Rasch item difficulty estimates are continuous logit measures, this study discretised them into three difficulty levels to support operational decision-making within the Error-Bound Pareto Frontier Framework. The aim was not to replace continuous psychometric interpretation, but to translate calibrated difficulty into deployment-relevant strata that can guide model assignment, cost-latency optimisation, and human-in-the-loop decisions. Continuous modelling of LLM accuracy against item difficulty, such as through item characteristic curve-style analyses, would provide a complementary perspective. However, the present framework prioritised interpretable difficulty bands because institutional deployment decisions are typically made at actionable levels of risk and complexity rather than at each individual logit value.

3.2 Rasch Analysis

Prior to LLM evaluation, Rasch analysis was conducted using RUMM2030 to calibrate MCQs for each course. Item difficulty parameters were estimated on a common logit scale, and item fit was assessed based on fit residuals. The Person Separation Index (PSI) ranged from .75 to .92 for the three courses, indicating acceptable to good reliability across the courses. The person and item separation indices ranged from .77 to .94 across the courses, indicating adequate to good separation of participants and item difficulties. In addition, all items demonstrated acceptable fit to the Rasch model, with fit residuals falling within the recommended threshold of −2.5 to +2.5 (e.g., Andrich & Marais, 2019), indicating no substantive item misfit. These

results supported the use of the calibrated item set for subsequent difficulty-aware LLM evaluation.

3.3 Large Language Models

The accuracy and performance of five LLMs: 4o, 4o-mini, o3-mini, gpt5-nano and DeepSeek-V3 on MCQs were evaluated. The OpenAI API was employed to run the 4o, 4o-mini, o3-mini and gpt5-nano models, while the DeepSeek API was utilised to run DeepSeek-V3. All models were run with default API parameters, with temperature set to 0 where applicable. The reasoning efforts of o3-mini and gpt5-nano were set at medium.

3.4 Prompt Design

To encourage the generation of concise response, for each MCQ, we provided every model with a standardised prompt. Additionally, all options for each MCQ were randomly arranged with only one option being correct (i.e., Mathematics, four options; Psychology and Statistics, three options). An example of prompt format used is shown in Figure 1.

Figure 1

An Example of Prompt Used in this Study

```
You are given a question in the context below consisting of a question and three options (a, b, c). Your task is to answer the multiple-choice question and choose the correct answer. Do not return the full text answer, just the alphabet option.
Context: {'question_id': '521', 'q.text': 'What is the main significance of the five-factor trait theory?', 'a': 'It proposes a reasonable number of well-defined personality factors that lend themselves to further research.', 'b': 'It further refines the personality factors identified in other trait theories.', 'c': 'Its five factors may represent universal personality components in both humans and other species.'}
Return your answer using the following JSON format:
{
  'question_id': <question_id>,
  'model_answer': <a or b or c>
}
```

3.5 Model Evaluation Metrics

To evaluate the LLMs, data were collected and analysed based on the following three primary criteria for each model and course:

Accuracy: The proportion of correct answers calculated overall and, critically, per difficulty level.

Cost: Operationalised as the average monetary cost (USD) incurred per MCQ request, calculated by summing input and output token usage based on model-specific pricing. To improve interpretability and support meaningful comparisons, all costs were scaled to reflect the cost per 100 questions. This normalisation was applied to prevent extremely small unit costs (e.g., <$0.001) from obscuring meaningful cost differences across models and limiting their interpretability for practical decision-making.

Latency: Captured as the response time (in seconds) for a single MCQ request. In line with the Human-Computer Interaction (HCI) literature (Seow, 2008; Nielsen, 1993), the 90th percentile (p90) latency was reported as a proxy for cognitive usability in real-time applications.

3.6 Error-Bound Pareto Frontier

Rather than selecting the single best model based on average performance, we employed the Error-Bound Pareto Frontier as an extension of Pareto optimization (Pareto, 1906/1971); Deb, 2001) as a decision-making framework tailored to specific educational deployment contexts. Under this approach, a strict error threshold (e.g., 5% error) was enforced as a primary constraint at each difficulty level (i.e., Levels 1-3) within a course. Models that failed to meet this threshold at a given difficulty level were excluded from assessing items at that level. The framework allowed models to switch across difficulty levels, assigning different models depending on which satisfied the error constraint most effectively (Erdil, 2025). Among the qualifying models at each difficulty level, Pareto optimisation was applied to select the best option based on cost or latency, depending on the deployment scenario.

The 5% error bound, operationalised as a ≥95% accuracy threshold, was used as a conservative deployment guardrail rather than as a universal standard. In educational assessment contexts, especially where generated or answered MCQs may influence learning progression or item-bank quality, even small error rates can create downstream risks if they are scaled across large numbers of items. The threshold therefore represents a practical institutional tolerance for low-error automation. Importantly, the framework is threshold-flexible: institutions may adopt stricter or more relaxed error bounds depending on assessment stakes, availability of human review, resource constraints, and intended use.

While accuracy remains the primary criterion in defining the Error-Bound Pareto Frontier, cost and latency are not merely secondary optimisation parameters but can be incorporated into the error-bound framework. Drawing from HCI literature (Seow, 2008; Nielsen, 1993), latency thresholds can be operationalised as boundary conditions where response times beyond 10 seconds are not just inconvenient, but potentially disruptive to user engagement and learning flow. In this extended formulation, models that consistently exceed this latency threshold may be excluded from real-time applications, regardless of their accuracy. Similarly, cost ceilings, defined by institutional or deployment-specific constraints, can function as hard filters rather than soft preferences. Thus, in practice, the error-bound frontier can be multi-dimensional, where a model must satisfy not only an accuracy threshold but also fall within acceptable cost and latency bounds. This enables a more holistic and deployable framework, allowing stakeholders to systematically exclude models that breach cognitive usability thresholds or economic feasibility constraints.

4 Results and Analysis

This section presents the results on LLM accuracy, and cost and latency based on the Rasch-calibrated MCQs used in this study, followed by the corresponding outcomes of the Error-Bound Pareto Frontier Framework. These LLM results served as inputs to the framework.

4.1 Overall Model Accuracy

Table 1 shows the overall model accuracy for Mathematics, Psychology and Statistic courses, with results aggregated into three main subject areas. These results emphasise the importance of selecting models based on their strengths to optimise performance in specific subject areas.

Table 1

Overall accuracy by Model and Course

Model / Course	1 n=480	2 n=400	3 n=140
		% accuracy	
4o	60.7	89.3	81.4
4o-mini	46.0	84.3	34.3
o3-mini	97.5	90.5	87.9
gpt5-nano	97.7	89.0	88.6
DeepSeek-V3	82.1	88.3	83.6

gpt5-nano: This model performed exceptionally well across all three courses, scoring 97.7% in Mathematics, 89.0% in Psychology, and 88.6% in Statistics. This positioned gpt5-nano as a top-tier alternative for applications requiring high accuracy with stable cross-domain reliability, and it was the top performer in Mathematics and Statistics in this evaluation.

o3-mini: This model performed exceptionally well across all three courses, scoring 97.5% in Mathematics, 90.5% in Psychology, and 87.9% in Statistics. It was the top-performing model in Psychology and demonstrated consistently high accuracy across the three subject areas.

4o: The 4o model performed strongly in Psychology (89.3%) and reasonably well in Statistics (81.4%), but it underperformed in Mathematics (60.7%). The findings were consistent with the results from the Massive Multitask Language Understanding (MMLU). OpenAI (2024) reported a score of 88.7% on the MMLU benchmark which includes tasks related to psychology. On the contrary, the performance of 4o in Mathematics courses was relatively lower. This disparity reflects the strengths of 4o in language comprehension over complex mathematical reasoning. As reported by OpenAI (2024), 4o solved only 12% of mathematics problems based on the 2024 AIME benchmark, underscoring its limitations in mathematical problem-solving.

4o-mini: While 4o-mini achieved moderate accuracy in Psychology (84.3%), its lower scores in Mathematics (46.0%) and especially Statistics (34.3%) suggest limitations in managing complex reasoning tasks. These results are expected given the reduced computational capacity of the mini variant, which may lack the depth required for high-stakes analytical assessments.

DeepSeek-V3: This model provided balanced performance across the board: 82.1% in Mathematics, 88.3% in Psychology, and 83.6% in Statistics. These results reflect its positioning as a versatile model, confirmed by external benchmark results with 59.4% on AIME 2024 and 81.2% on the MMLU (DeepSeek, 2025).

4.2 Model Accuracy by Difficulty Levels

Tables 2, 3 and 4 present the model accuracy across three difficulty levels (Level 1 to Level 3) derived from Rasch-calibrated item difficulty estimates. These levels reflected increasing item challenges, offering a granular view of how models handle low to high difficulty MCQs across subjects.

Mathematics (Course 1): As shown in Table 2, o3-mini and gpt5-nano consistently outperformed the other models across all three difficulty levels. Both achieved perfect accuracy at Level 1 (100.0%), demonstrating strong mastery of foundational mathematical content. At Level 2, performance remained very high (gpt5-nano: 97.4%, o3-mini: 96.7%), and at Level 3 both models continued to lead (o3-mini: 97.3%, gpt5-nano: 96.4%), indicating robust reasoning even as item difficulty increased.

Table 2

Accuracy for Mathematics by Difficulty Level

Difficulty Level	Total MCQs	4o	4o-mini	o3-mini	DeepSeek-V3	gpt5-nano
		% accuracies				
1	101	79.2	63.4	100.0	92.1	100.0
2	269	52.7	45.7	96.7	82.9	97.4
3	110	63.2	31.0	97.3	70.9	96.4

DeepSeek-V3 showed solid performance at Level 1 (92.1%) but declined as difficulty increased (82.9% at Level 2 and 70.9% at Level 3), suggesting reduced reliability for more challenging content.

In contrast, 4o exhibited an uneven pattern - moderate performance at Level 1 (79.2%), a sharp drop at Level 2 (52.7%), and a partial rebound at Level 3 (63.2%), indicating instability across difficulty tiers. 4o-mini remained the weakest model, with accuracy decreasing notably at higher complexity (63.4% → 45.7% → 31.0% from Level 1 to Level 3).

Psychology (Course 2): In Table 3, most models performed near ceiling at Level 1, indicating strong capability for foundational psychology questions. 4o achieved the highest Level 1 accuracy (99.4%), followed closely by o3-mini and DeepSeek-V3 (98.7% each), with gpt5-nano at 97.5% and 4o-mini at 96.9%.

Table 3

Accuracy for Psychology by Difficulty Level

Difficulty Level	Total MCQs	4o	4o-mini	o3-mini	DeepSeek-V3	gpt5-nano
		% accuracies				
1	159	99.4	96.9	98.7	98.7	97.5
2	187	87.2	81.3	90.4	85.6	89.3
3	54	66.7	57.4	66.7	66.7	63.0

At Level 2, clearer separation emerged: o3-mini led (90.4%), followed by gpt5-nano (89.3%), 4o (87.2%), and DeepSeek-V3 (85.6%), while 4o-mini lagged further behind (81.3%).

At Level 3, performance decreased at the higher Rasch-calibrated difficulty level. 4o, o3-mini, and DeepSeek-V3 converged at 66.7%, while gpt5-nano scored 63.0% and

4o-mini recorded the lowest accuracy (57.4%). This pattern suggests that items at higher difficulty may require nuanced reasoning beyond current model capabilities.

Table 4

Accuracy for Statistics by Difficulty Level

Difficulty Level	Total MCQs	4o	4o-mini	o3-mini	DeepSeek-V3	gpt5-nano
		% accuracies				
1	32	96.9	37.5	100.0	100.0	100.0
2	70	92.9	37.1	98.6	91.4	94.3
3	38	47.4	26.3	57.9	55.3	68.4

Statistics (Course 3): Table 4 shows the most pronounced separation between models, especially at higher Difficulty Levels. At Level 1, o3-mini, DeepSeek-V3, and gpt5-nano achieved perfect accuracy (100.0%), while 4o also performed strongly (96.9%). In contrast, 4o-mini performed markedly lower (37.5%), indicating major limitations even at the foundational level in statistics.

At Level 2, o3-mini remained the strongest (98.6%), followed by gpt5-nano (94.3%), 4o (92.9%), and DeepSeek-V3 (91.4%), while 4o-mini again remained low (37.1%).

4.3 Rasch-Consistent Performance Patterns

The difficulty-stratified results also provide insight into whether LLM performance followed expectations consistent with Rasch-calibrated difficulty. Overall, the results showed broadly monotonic, Guttman-like features, with accuracy generally decreasing as Rasch-calibrated difficulty increased. This suggests that the Rasch-calibrated difficulty levels captured aspects of task complexity that were relevant to LLM performance. However, the pattern was not perfectly consistent across all model-subject combinations. For example, in Mathematics, 4o performed lower at Level 2 than at Level 3, while o3-mini showed a slight increase from Level 2 to Level 3. These exceptions suggest that LLM performance may also be influenced by item features other than human-calibrated difficulty, such as wording, symbolic reasoning demands, distractor plausibility, or domain-specific familiarity. Future work could examine model-level response patterns more directly, including whether LLM accuracy follows monotonic trends across continuous Rasch logit difficulty.

4.4 Cost and Time Efficiency

The cost per 100 MCQs was calculated by dividing total token cost by the number of questions, then scaling to 100 requests for cross-model comparison. Table 5 shows that 4o-mini and DeepSeek-V3 emerged as the most cost-efficient models, with costs as low as $0.002–$0.006 per 100 questions, depending on the subject. This makes them

especially suitable for high-volume or budget-constrained deployments, such as large classrooms or open-access platforms. For gpt5-nano, the model strikes a favourable balance between cost and accuracy, with costs ranging from $0.013 (Statistics) to $0.018 (Psychology) which is approximately 3.4–6.5 times more expensive than the cheapest models, but more accurate.

Model 4o presented moderate costs from $0.036 to $0.090 per 100 MCQs, higher than 4o-mini and DeepSeek-V3, yet faster and more accurate in select verbal domains.

Model o3-mini was the most expensive model, with costs ranging from $0.076 (Statistics) to $0.105 (Psychology) per 100 MCQs. While its accuracy is among the highest, the cost profile poses a challenge for large-scale or long-term deployments.

Table 5

Cost per 100 MCQ Request (One Question and Answer) by Course and Model

Model / Course	Mathematics	Psychology	Statistics
	$ USD / 100 question-answer		
4o	0.090	0.068	0.036
4o-mini	0.005	0.004	0.002
o3-mini	0.093	0.105	0.076
gpt5-nano	0.017	0.018	0.013
DeepSeek-V3	0.005	0.006	0.003

Table 6

p90 Time Taken per MCQ Request (One Question and Answer) by Course and Model

Model / Course	Mathematics	Psychology	Statistics
	seconds		
4o	1.8	1.7	1.6
4o-mini	2.0	2.0	1.5
o3-mini	3.8	4.4	7.4
gpt5-nano	6.4	10.2	6.8
DeepSeek-V3	2.1	2.1	2.2

Table 6 reports the p90 latency, which captures the time within which 90% of responses were returned. According to HCI literature (Seow, 2008; Nielsen, 1993), human perception of system latency falls into four categories: latencies under 0.1 seconds feel instantaneous; between 0.1 and 1 second are perceptible but maintain a fluent user experience; 1 to 10 seconds are noticeable delays that may interrupt

cognitive flow; and beyond 10 seconds, latency becomes disruptive to interaction. In this context, 4o and 4o-mini demonstrated the fastest response times, consistently at or below 2 seconds across all courses, placing them in the "noticeable but tolerable" range for online interactive educational applications. DeepSeek-V3 also showed stable latency around 2.1 to 2.2 seconds, remaining within acceptable bounds for synchronous use. In contrast, o3-mini exhibited higher latency, particularly in Statistics (7.4 seconds) and Psychology (4.4 seconds), approaching the upper threshold where delays may disrupt user attention. gpt5-nano showed the highest latency in Mathematics and Psychology, with Psychology responses exceeding 10 seconds, placing it in the disruptive category. This makes it more appropriate for asynchronous workflows such as batch scoring or delayed content review, rather than real-time educational interactions.

4.5 Error-Bound Pareto Frontier

Figures 2 to 4 present the Pareto Frontier for Mathematics across levels. In Mathematics, o3-mini and gpt5-nano were the only models that consistently satisfy the ≥95% accuracy constraint across the three levels of difficulty. At Level 1, both models achieved 100.0% accuracy. At Level 2, both remained above threshold (o3-mini: 96.7%, gpt5-nano: 97.4%), and at Level 3 both still cleared the bound (o3-mini: 97.3%, gpt5-nano: 96.4%). Given that both models met the error bound across all levels, selection was primarily a deployment trade-off: gpt5-nano was the cost-optimal option in Mathematics ($0.017 per 100 requests) relative to o3-mini ($0.093), while o3-mini was the latency-optimal option (3.8s vs 6.4s). Thus, IHLs could consider maintaining strict item generation quality across mathematical difficulty levels while optimizing either for budget (gpt5-nano) or responsiveness (o3-mini).

Figure 2

Pareto Frontier for Mathematics at Difficulty Level 1

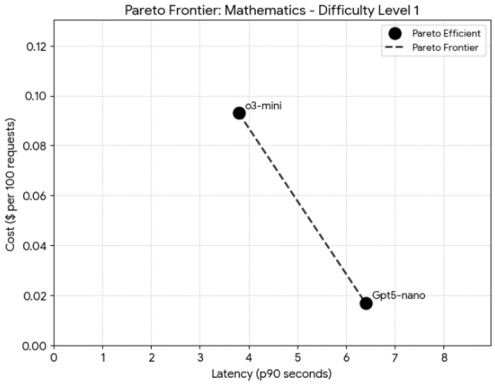


Figure 3

Pareto Frontier for Mathematics at Difficulty Level 2

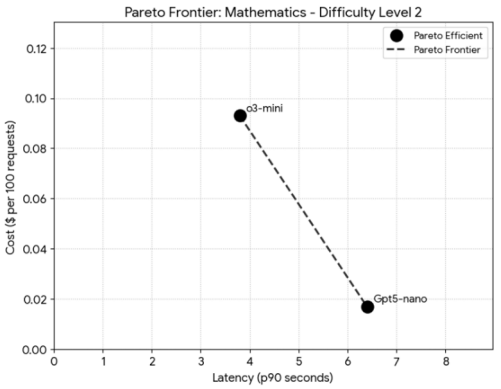


Figure 4

Pareto Frontier for Mathematics at Difficulty Level 3

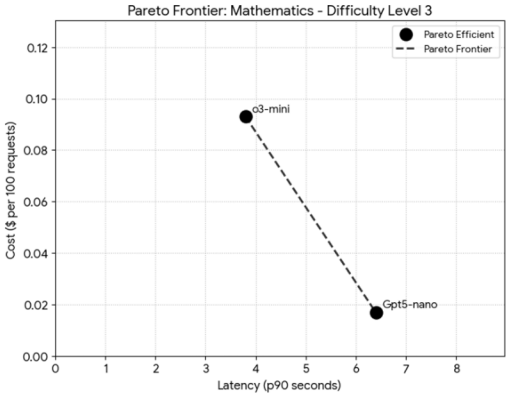


Figure 5 shows that the error-bound frontier existed only at Level 1 in Psychology. At Level 1, all models satisfied the ≥95% accuracy threshold (96.9%–99.4%), enabling automation under strict tolerance for foundational psychology items. Under Pareto selection for Level 1, 4o-mini was the cost-minimiser ($0.004 per 100 requests) and 4o was the latency-minimiser (1.7s). Models such as DeepSeek-V3, o3-mini, and gpt5-nano were not Pareto-efficient in this tier. DeepSeek-V3 was strictly dominated by 4o-mini (higher cost and latency), while the others were either more expensive and slower (o3-mini), or substantially slower while not being the cheapest option (gpt5-nano). However, for Level 2 and Level 3, no model met the 5% error tolerance (best performance is 90.4% at Level 2; and 66.7% at Level 3). Consequently, Psychology

was a partial-automation course under strict error bounds: Level 1 could be automated with Pareto-optimised model selection, while Levels 2 and 3 required human review because no evaluated model met the predefined accuracy threshold.

Figure 5

Pareto Frontier for Psychology at Difficulty Level 1

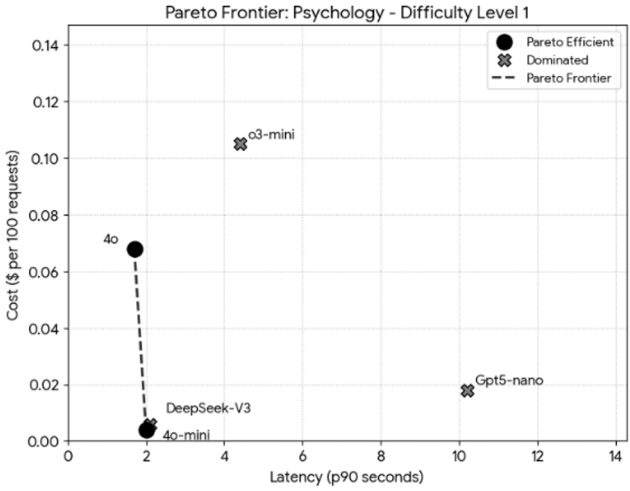


Figure 6

Pareto Frontier for Statistics at Difficulty Level 1

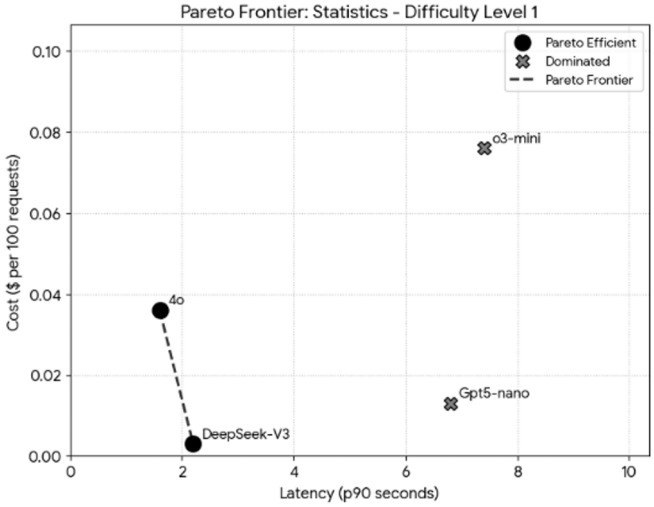
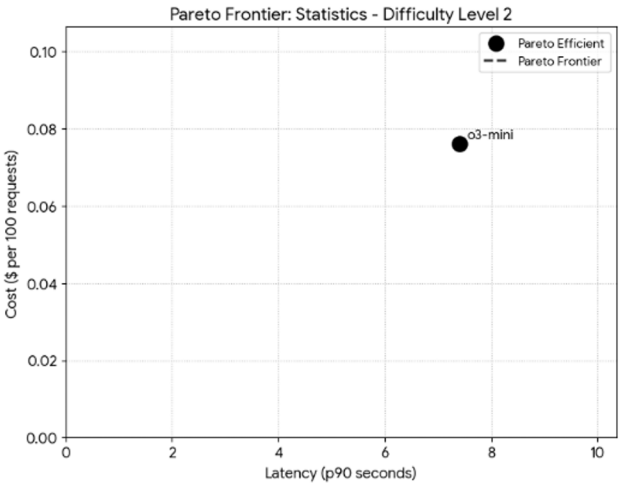


Figure 7

Pareto Frontier for Statistics at Difficulty Level 2



Figures 6 and 7 present the Pareto Frontier for Statistics at Levels 1 and 2. In Statistics, the feasible set shrunk sharply as complexity increased. At Level 1, four models met the ≥95% accuracy threshold (4o: 96.9%, o3-mini: 100.0%, DeepSeek-V3: 100.0%, gpt5-nano: 100.0%); notably, 4o-mini was excluded from the feasible set (Figure 6) due to accuracy failure, despite its low cost. Within this feasible set, the efficient frontier was defined by DeepSeek-V3 (the cost-minimizer, $0.003, 2.2s) and 4o (the latency-minimizer, 1.6s, $0.036). Notably, o3-mini was strictly dominated by 4o at Level 1 (higher cost and higher latency), and gpt5-nano was strictly dominated by DeepSeek-V3 (higher cost and higher latency), making them Pareto-inefficient choices despite perfect accuracy. At Level 2, the frontier collapses to a single feasible model: only o3-mini (Figure 7) met the ≥95% threshold (98.6%). This eliminated cost/latency trade-offs under a strict error bound, and implied that Level 2 automation, if permitted, came with substantial resource implications ($0.076 per 100, 7.4s p90). At Level 3, no model reached the threshold (best is gpt5-nano at 68.4%), creating a "null frontier" where strict automation was not permissible and human review remained necessary.

5 Limitations and Future Research

This study has several limitations. First, the analysis was based on a single institutional dataset of 1,020 MCQs from Mathematics, Psychology, and Statistics. Although the item bank was Rasch-calibrated, the findings may not generalise to other institutions, disciplines, item formats, or assessment levels. Second, the evaluation was limited to five LLMs available at the time of testing. Given rapid model turnover, future work should treat the framework as a repeatable evaluation procedure rather than as a

fixed ranking of specific models. Third, the evaluation used a standardised zero-shot prompt, with default API settings and medium reasoning effort for selected reasoning models. Alternative prompting strategies, including few-shot prompting, chain-of-thought-style prompting, or different reasoning-effort settings, may interact with item difficulty and alter the frontier. Fourth, while the framework evaluates model accuracy in answering calibrated MCQs, it does not include direct human validation of LLM-generated items. Future studies should include subject-matter expert review of generated questions, answer keys, distractors, cognitive level, and alignment with learning outcomes.

Future research should extend this work by evaluating additional disciplines, larger and more diverse calibrated item banks, and newer LLMs as they become available. Continuous analyses of accuracy against Rasch logit difficulty may also complement the current difficulty-band approach. In addition, future studies should examine how human-in-the-loop review can be systematically incorporated at higher difficulty levels, particularly where no model meets the required error bound.

6 Conclusion

This paper proposes a version-robust Error-Bound Pareto Frontier Framework for selecting LLMs in automated MCQ workflows, grounded in Rasch-calibrated item difficulty. By applying strict accuracy guardrails before optimising cost and latency, the framework holds promise as an approach for navigating rapid model churn while supporting scalable and pedagogically informed deployments.

Ultimately, the framework may support durable, context-sensitive LLM selection through Rasch-calibrated difficulty-level model assignment, balancing instructional reliability with operational constraints. However, higher-difficulty items, especially where no model meets the required error bound, should remain subject to human-in-the-loop review. Such review should include subject-matter expert verification of content accuracy, answer keys, distractor quality, alignment with learning outcomes, and appropriateness for the intended cognitive level.

Author Contributions. Conceptulisation, Che Yee Lye; methodology, Hariz Liew; analysis, Hariz Liew, Che Yee Lye; writing-original draft preparation: Hariz Liew; writing-reviewing and editing, Che Yee Lye. Both authors have read and agreed to the published version of the manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory: Measuring in the educational, social and health sciences. Springer Nature Singapore. https://doi.org/10.1007/978-981-13-7496-8
2. Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40. https://doi.org/10.1177/0265532220927487
3. Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023). Distractor generation for multiple-choice questions with predictive prompting and large language models. arXiv preprint arXiv: 2307.16338.
4. Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms*. John Wiley & Sons.
5. DeepSeek. (2025, March 25). DeepSeek-V3-0324 release. https://api-docs.deepseek.com/news/news250325
6. Erdil, E. (2025). Inference economics of language models. arXiv preprint arXiv: 2506.04645
7. Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., & Fernandez-Llorca, D. (2025). Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. arXiv preprint arXiv: 2502.06559
8. Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6). https://doi.org/10.3102/0034654317726529
9. Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3). https://doi.org/10.1207/S15324818AME1503_5
10. Hwang, K., Wang, K., Alomair, M., Choa, FS., & Chen, L. K. (2024). Towards automated multiple choice question generation and evaluation: Aligning with Bloom's Taxonomy. In A. M. Olney, IA., Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds), *Artificial Intelligence in Education, AIED 2024. Lecture Notes in Computer Science, 14830*, Springer, Cham. https://doi.org/10.1007/978-3-031-64299-9_35
11. Imhof, C., Bergamin, P., & McGarrity, S. (2020). Implementation of adaptive learning systems: Current state and potential. In P. Isaias, D. G. Sampson, & D. Ifenthaler (Eds.), *Online teaching and learning in higher education: Cognition and exploratory learning in the digital age*. Springer, Cham. https://doi.org/10.1007/978-3-030-48190-2_6
12. Jenks, G. F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography,* 7, 186–190.

13. Khosravi, H., Sadiq, S., & Gasevic, D. (2020). Development and adoption of an adaptive learning system: Reflections and lessons learned. SIGCSE'20: *Proceedings of the 51st ACM Technical Symposium on Computer Science Education.* https://doi.org/10.1145/3328778.3366900
14. Kumar, A. P., Nayak, A., Shenoy, M. K., Goyal, S., & Chaitanya. (2023). A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, 225, 120022. https://doi.org/10.1016/j.eswa.2023.120022
15. Li, W., Li, L., Xiang, T., Liu, X., Deng, W. & Garcia, N. (2024). Can multiple-choice questions really be useful in detecting the abilities of LLMs? *LREC-COLING*. https://doi.org/10.48550/arXiv.2403.17752
16. Moss, E. (2001). Multiple choice questions: Their value as an assessment tool. *Curr Opin Anaesthesiol, 14*(6). https://doi.org/10.1097/00001503-200112000-00011
17. Mucciaccia, S., Paixao, T., M., Mutz, F., De Souza, A. F., Badue, C. S., & Oliveira-Santos, T. (2025). Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. *Proceedings of the 31st International Conference on Computational Linguistics*. https://aclanthology.org/2025.coling-main.154
18. Nielsen, J. (1993). *Usability engineering*. Academic Press.
19. Nikolovski, V., Trajanov, D., & Chorbev, I. (2025). Advancing AI in higher education: A comparative study of large language model-based agents for exam question generation, improvement and evaluation. *Algorithms, 18*(144). https://doi.org/10.3390/a18030144
20. Oc, Y., & Hassen, H. (2025). Comparing the effectiveness of multiple-answer and single-answer multiple-choice questions in assessing student learning. *Marketing Education Review, 35*(1). https://doi.org/10.1080/10528008.2024.2417106
21. OpenAI. (2024, September 12). OpenAI o1-mini: Advancing cost-efficient reasoning. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning
22. OpenAI. (2024, September 12). Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms
23. Pareto, V. (1971). *Manual of political economy* (A. S. Schwier, Trans.). Augustus M. Kelley. (Original work published 1906)
24. Ramesh, A., Agarwal, A., Doughty, J. A., Ramaneti, K., Savelka, J., & Sakr, M. (2024). A benchmark for testing the capabilities of LLMs in assessing the quality of multiple-choice questions in introductory programming education. *SIGCSE Virtual 2024*. https://doi.org/10.1145/3649165.3690123
25. Raina, V., & Gales, M. (2022). *Multiple-choice question generation: Towards an automated assessment framework*. arXiv preprint arXiv: 2209.11830
26. Rodriguez, P., Barrow, J., Hoyle, A., Lalor, J. P., Jia, R., & Boyd-Graber, J. (2021). Evaluation examples are not equally informative: How should that change NLP leaderboards? *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processings, 1, 4486–4503. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.346
27. Savelka, J., Agarwal, A., Bogart, C., & Sakr, M. (2023). Large language models (GPT) struggle to answer multiple-choice questions about code. *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023)*, 2. https://doi.org/10.5220/0011996900003470
28. Seow, S. C. (2008). *Designing and engineering time: The psychology of time perception in software*. Addison-Wesley Professional.
29. Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approach. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.00825
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv: 1706.03762
31. Wang, X., Jiang, L., Hernandez-Oralic, J., Stillwell, D., Sun, L., Luo, F., & Xie, X. (2023). Evaluating general-purpose AI with psychometrics. arXiv preprint arXiv: 2310.16379
32. Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.
33. Ye, H., Jin, J., Xie, Y., Zhang, X., & Song, G. (2025). Large language model psychometrics: A systematic review of evaluation, validation and enhancement. arXiv preprint arXiv: 2505.08245

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

