Automated item evaluation: Predicting item acceptance and rejection using LLM-generated critiques

Hotaka Maeda [1] Yikai Lu [2]
[1]Smarter Balanced, University of California-Santa Cruz
[2]Department of Educational Psychology, University of Minnesota-Twin Cities *

Abstract

Automated item evaluation (AIE) refers to the use of computational methods to assess item quality without requiring manual expert review or field testing of the items under evaluation. We aimed to build a near-comprehensive AIE model by predicting item acceptance and rejection from item text using historical rejection data from a large-scale standardized testing program. The dataset contained 52,759 English language arts (ELA) and mathematics items with 34% permanently rejected from future operational use. Rejection reasons included poor psychometric properties, content issues, bias and sensitivity concerns, and non-content issues. We fine-tuned a DeBERTaV3-large classifier on raw item text, a second DeBERTa classifier on Qwen3-generated item critiques, and a fusion model combining representations from both. The fusion model achieved the strongest overall performance (Accuracy = .75, F1 = .64, AUC= .80, Sensitivity = .64, Specificity = .81). Prediction for math (F1 = .73, AUC = .86) was considerably more accurate than ELA (F1 = .51, AUC = .72). Lowering the decision threshold from .5 to .25 raised average sensitivity for ELA and math to .88 and .91, while reducing specificity to .31 and .56, respectively, which may be preferable in automated item generation contexts where generating items is cheaper than evaluating them. Incorporating item critiques alongside raw item text improved performance across most rejection reasons. The model assigned higher rejection probabilities to more difficult items. However, the fusion model struggled to identify items flagged for bias, sensitivity, fairness, or accessibility, especially for ELA. These findings suggest that text-based AIE is feasible in some areas and may offer a practical tool for reducing the burden of manual review and field testing, while also underscoring the importance of human review for items with fairness concerns.

Keywords: natural language processing, Transformers, artificial intelligence, item difficulty prediction, item response theory

1 Introduction

Evaluating the quality of new items is expensive and time-consuming, requiring content review and field testing. This issue has grown since the emergence of automated item generation (AIG) using large language models (LLMs). AIG review papers consistently urge for methods that can quickly filter poor quality items (Oluoke et al., 2026; Circi et al., 2023; Falcão et al., 2022; Tan et al., 2025). Relevant literature has focused on using LLMs, natural language processing (NLP), or machine learning to predict individual qualities of items, including item difficulty (AlKhuzaey et al., 2023; Benedetto et al., 2023), content alignment with the assessment blueprint (Fu et al., 2025), or differential item functioning (Maeda and Lu, 2025). However, prior work has not addressed these concerns as a unified class of issues related to item quality. To capture this broader problem space, we use the term *automated item evaluation* (AIE; Yu and Burke, 2026), extending its prior use from a much narrower scope. In this study, AIE refers to the use of computational methods to automatically assess item quality without requiring manual expert review or field testing of the items under evaluation.

*Correspondence concerning this article should be addressed to Hotaka Maeda, Smarter Balanced, University of California-Santa Cruz, 1156 High St, Santa Cruz, CA 95064. E-mail: hotaka.maeda@smarterbalanced.org

The difficulty of AIE stems from the fact that there are countless reasons that items can be unsuited for operational use. The Standards for Educational and Psychological Testing (AERA et al., 2014) articulate the broader requirements that define a valid, reliable, and fair item. Also, Haladyna and Downing (1989) cataloged 43 item-writing rules covering stem construction, option formatting, correct answer selection, and distractor quality. Training and deploying a separate prediction model for each rule is unrealistic because model training requires large sample sizes and substantial computational resources. Moreover, treating rules separately fails to leverage the fact that items may exhibit multiple issues simultaneously and does not directly support the goal of rejecting the most undesirable items.

Therefore, a comprehensive AIE approach could be desirable. One possible approach is to use an existing item bank by treating operational items as "accepted" and items permanently retired due to irreparable issues as "rejected", then training a model to predict that outcome from item content alone. Accepted items had to have gone through rigorous content expert review and field testing, so this single label becomes an all-encompassing measure of item quality. Although this approach requires a large established item bank, it may be the single most comprehensive AIE method for both AIG and human-written items.

To our knowledge, no studies have predicted item acceptance and rejection as an AIE method. The purpose of this study is to train a transformer language model to classify items as accepted or rejected for operational use, using item status data from a large-scale standardized testing program. To enhance predictive power, we augment the data using item critiques generated by an LLM. To provide insight into model behavior beyond overall classification performance, we use historical item developer comments to evaluate classifier sensitivity by rejection reason. We also include a sentiment analysis to further understand how LLM critique sentiment is related to prediction model behavior.

2 Related Works

2.1 (Automated) Item Evaluation

Item evaluation is considered as an important step during test development (AERA et al., 2014; Haladyna and Downing, 1989). Gorgun and Bulut (2025) summarized existing item evaluation approaches into three categories: metric-based evaluations, post-hoc analysis, and human evaluations. Metric-based approaches typically focus on comparing automatically generated items with a set of reference items (Gorgun and Bulut, 2025). Other metric-based evaluation approaches include NLP-based techniques, such as word count, TF-IDF, word embeddings, cosine similarity for identifying enemy items, and measures of readability and text complexity (Yu and Burke, 2026; Amini et al., 2025). However, despite their potential efficiency, metric-based approaches are limited by their reliance on reference items. This dependence can significantly restrict the content domain of generated items, as valid items may be incorrectly penalized simply because they *appear* different from reference items at the level of linguistic features (Gorgun and Bulut, 2025). Post-hoc analysis approaches typically involve field testing, in which real item response data are collected from a representative sample of examinees to assess item quality. The most commonly examined aspects are statistical item properties, such as item difficulty prediction (AlKhuzaey et al., 2023; Benedetto et al., 2023) and item discrimination (e.g., Maeda and Lu, 2026; Han et al., 2025; Yaneva et al., 2020). Less commonly examined areas include differential item functioning (Maeda and Lu, 2025). However, this approach can be costly and time-consuming, as it requires substantial resources to collect response data and obtain reliable statistical estimates. Finally, human evaluators can identify problems in items that may be difficult to detect using computational or statistical approaches. However, human evaluation can be subjective and time-consuming, especially when many items must be reviewed. Another limitation is that evaluators need to be familiar with the relevant content domain, which can make it difficult to identify suitable experts for the evaluation process.

The term "automated item evaluation" or a similar term has been used occasionally typically in the AIG literature (Gorgun and Bulut, 2025; Prentzas

and Binopoulou, 2025; Yu and Burke, 2026; Shin and Gierl, 2024; Wang et al., 2025). Despite its sparse usage and variation in scope in the past literature, definitions of AIE appear to share a common theme: the use of computational methods to automatically assess some facets of item quality without requiring manual expert review or field testing of the items under evaluation. Under this definition, only the first approach described by Gorgun and Bulut (2025) meets the criteria for AIE, as it eliminates the need to rely on field testing or human evaluators.

However, although not proposed by Gorgun and Bulut (2025) themselves, the approach they described could be considered a fourth category: LLM-based evaluation, which is another AIE approach. This approach is the closest precedent to our comprehensive evaluation framework, as it trained an LLM to identify low-quality AIG items using human review labels. However, this approach is severely limited because it provides only a coarse-grained evaluation based on human review labels, classifying items simply as *good* or *bad*. Note that, as Pelánek et al. (2022) argues, classifying items as *good* or *bad* may be a fundamentally ill-defined problem, as items that warrant the attention of content creators depend on the specific context and are at least partly subjective. Furthermore, items may be problematic for many different reasons, including linguistic or psychometric characteristics that sometimes cannot be captured by human evaluators and vice versa. Therefore, it is important for us to consider different aspects of item quality at once.

It is worth highlighting that LLM-based evaluation has made possible the automatic assessment of item-content alignment with assessment blueprints, a less commonly examined area of AIE (Fu et al., 2025). This is an important consideration in exam design in its own right, as it directly affects content validity. This suggests that LLM-based evaluation could shed light on aspects of items that were difficult to evaluate automatically prior to LLMs.

The major limitation of these previous approaches is that they tend to focus on only one aspect of item quality at a time. In practice, however, different dimensions of item quality may be interrelated. For example, a difficult mathematics item may also contain difficult keywords, and such relationships cannot be fully leveraged when these issues are evaluated separately. Furthermore, relying solely on computational or statistical methods may overlook problems that can only be identified through human judgment. For these reasons, it is preferable to construct a model that can evaluate multiple dimensions of item quality simultaneously. For example, Pelánek et al. (2022) proposed a system that uses an outlier detection mechanism to identify computationally or statistically deviant items, thereby flagging potentially *bad* items. Similarly, Yaneva et al. (2020) predicted whether MCQs has acceptable item difficulty and discrimination in a high-stakes medical exam, framing the problem as binary classification using linguistic features and word embeddings extracted from item text. However, the linguistic features both studies considered were mostly text complexity measures, which by themselves cannot account for the semantic content of items, something that LLMs can leverage. On the other hand, our hybrid approach can account for these different aspects of item quality by using a dataset that includes rejected items with a broad range of documented reasons for rejection. Rather than relying solely on human evaluators or linguistic features, this approach fine-tunes an LLM to distinguish between acceptable and unacceptable items while implicitly learning from multiple sources of item-quality evidence by taking into account the content of items.

2.2 Transformer language models for text classification

The rise of Transformer language models increased demand for AIE research by making large-scale AIG feasible, while simultaneously expanding the computational tools useful for AIE. Traditionally, prediction from text in the assessment context has relied on expert judgment (Wauters et al., 2012), syntactic features such as word count and term frequency (Benedetto et al., 2020), or semantic features like word embeddings (Hsu et al., 2018). More recently, Transformer-based language models have substantially improved prediction accuracy (Li et al., 2025; Han et al., 2025; Maeda, 2025). Introduced by Vaswani et al. (2017), the Transformer architecture replaced recurrent networks with parallel attention mechanisms, enabling efficient train-

ing and scalable modeling of long-range dependencies. Models such as BERT (Devlin et al., 2019) are pre-trained on large corpora and fine-tuned for downstream tasks. Text is tokenized into embeddings and passed through encoder layers to produce contextually enriched representations.

In this study, we use DeBERTaV3-large (He et al., 2021). DeBERTa extends BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). DeBERTa outperforms earlier models by disentangling content and positional information within its attention mechanism, where each token is represented by separate content and position vectors, allowing the model to capture relationships between words more effectively. The V3 variant further incorporates replaced token detection, yielding a 304-million-parameter model with strong benchmark performance that even outperforms the newer ModernBERT in natural language understanding (Warner et al., 2024).

2.3 Using LLM Critiques for AIE

Recently, language models have been used to extract reasoning or rationales to augment a task-specific model (Henrichsen and Krebs, 2025; Hsieh et al., 2023; Scarlatos et al., 2025). For an example in the assessment context, GPT-4o model (OpenAI, 2024) was used to generate the reasoning steps required to reach each multiple choice item option, and the responses were inputted into the longformer encoder language model to predict item difficulty (Feng et al., 2025). In another example, LLMs were given the role of content experts to make absolute and pairwise item comparisons to estimate item difficulty (Kolesnikova et al., 2026).

We use a similar data augmentation technique in this study using Qwen3 (Qwen Team, 2025). The Qwen3 model family includes a series of compact dense variants spanning 0.6B, 1.7B, 4B, and 8B parameters, all released publicly under the Apache 2.0 license. These models follow the broader trend of developing capable small language models (SLMs) suited to resource-constrained settings, where low memory footprint and fast inference are primary constraints (Lu et al., 2025). A key capability distinguishing these models from earlier SLMs is the integration of a *thinking mode* and a *non-thinking mode* within a single model. Rather than maintaining separate chat and reasoning model variants, users can toggle between slow, multi-step reasoning and fast, context-driven response generation at inference time.

Empirical results reported in Qwen Team (2025) indicate that the Qwen3-8B, 4B, and 1.7B base models each outperform Qwen2.5 models of the next larger size class on standard benchmarks, reflecting substantial efficiency gains from the distillation-centric training approach.

3 Methods

3.1 Item Data

Included in the study were 52,759 items designed for English language arts (ELA) and mathematics state summative assessments for grades 3–11. Of these, 34% have been rejected, defined as permanently removed from future operational use (30% for ELA, 39% for math). Rejection can occur at any stage of item development, before or after field testing, or after some operational use. Rejection reasons span a range of categories including content and standards alignment issues, psychometric reasons, accessibility and sensitivity flags, or issues unrelated to the item content, such as data integrity problems (see Appendix A for full category descriptions).

Items that had not been rejected and had not yet been field tested were excluded from the study, as their operational suitability remains unknown. Items rejected prior to field testing were retained, as their unsuitability for operations is established regardless of psychometric properties. In total, 14% of items in the study are items rejected prior to field testing. The remaining 86% of items were field tested among grade-matched students across multiple states in the United States.

To prepare each item for modeling, item text was concatenated with a `\n` separator between prompts and any available answer options. Both selection-based (e.g., multiple choice, multiple select) and constructed response (e.g., short answer, extended response) items were included in the study. Any reading or listening passages were excluded from the item text data due to their extensive lengths. Any images, figures, or tables in the items were also excluded. Metadata were exluded as well, which included the scoring rubric and correct an-

swer keys. Data were partitioned randomly into approximately 80% training, 10% validation, and 10% test data. Item groups with multiple items associated with the same stimulus were always included in the same data group.

3.2 Prediction Approaches

Five prediction approaches were compared. We used a single NVIDIA A10G Tensor Core 24GB graphics processor for all model inferences and training.

3.2.1 Zero-Shot Baseline

Qwen3-0.6B (Qwen Team, 2025) was prompted to classify each item as "accept" or "reject" directly, without any task-specific fine-tuning (see Appendix B for prompt). The subject, grade level, and the aligned common core standards for the item were included in the prompt.

3.2.2 Raw Item Text Fine-Tuned Model

Item text was tokenized and passed into DeBERTaV3-large (He et al., 2021), fine-tuned as a binary classifier using Low-Rank Adaptation (LoRA; Hu et al. 2022; LoRA rank = 32, LoRA Alpha = 64, LoRA dropout = 0.05, dropout = 0.1, batch size = 8, max token length = 512, weight decay = .01, learning rate = 5e-5, epochs = 3). LoRA is a parameter-efficient fine-tuning method that inserts trainable low-rank matrices into selected layers while keeping the pretrained weights frozen, substantially reducing memory and computational costs. We added a binary classification head with a sigmoid function to convert logits into rejection probabilities, followed by cross-entropy loss (CEL) to minimize the distance between predicted and true labels.

There were approximately double the number of accepted items compared to rejected items. This class imbalance was addressed by incorporating class weights into the binary CEL. Weights were computed as inversely proportional to class frequency, such that the minority class received proportionally greater influence on the gradient updates during training. This approach preserves the full training sample while discouraging the model from defaulting to the majority class, effectively rebalancing the loss contribution of each class without altering the composition of the training data. By increasing the relative penalty for misclassifying rejected items, class weighting was expected to make the model more attentive to rejected items. We retained the state with the lowest validation CEL out of all epochs.

3.2.3 Item Critique Fine-Tuned Model

To augment the raw item text, Qwen3-0.6B (Qwen Team, 2025) with thinking mode enabled was prompted to generate a two sentence critique or praise of each item's quality. The subject, grade level, and the aligned common core standards for the item were included in the prompt. See Appendix C for the prompt and Appendix D for example critiques. The resulting critique text was then used as input to DeBERTaV3-large in place of the raw item text, with the same fine-tuning procedure as the text-only model (LoRA rank = 50, LoRA Alpha = 100, LoRA dropout = 0.05, dropout = 0.1, batch size = 8, max token length = 512, weight decay = .01, learning rate = 5e-5, epochs = 3).

3.2.4 Text + Critique Fusion Model

The fusion model combines representations from both DeBERTaV3-large models described above. The 1024-dimensional output of each model's final layer was extracted and concatenated. The DeBERTa weights were frozen, and only the fusion layers were trained. We added a 128-dimensional intermediate layer followed by a binary classification head with a sigmoid function and a weighted CEL (see Figure 1).

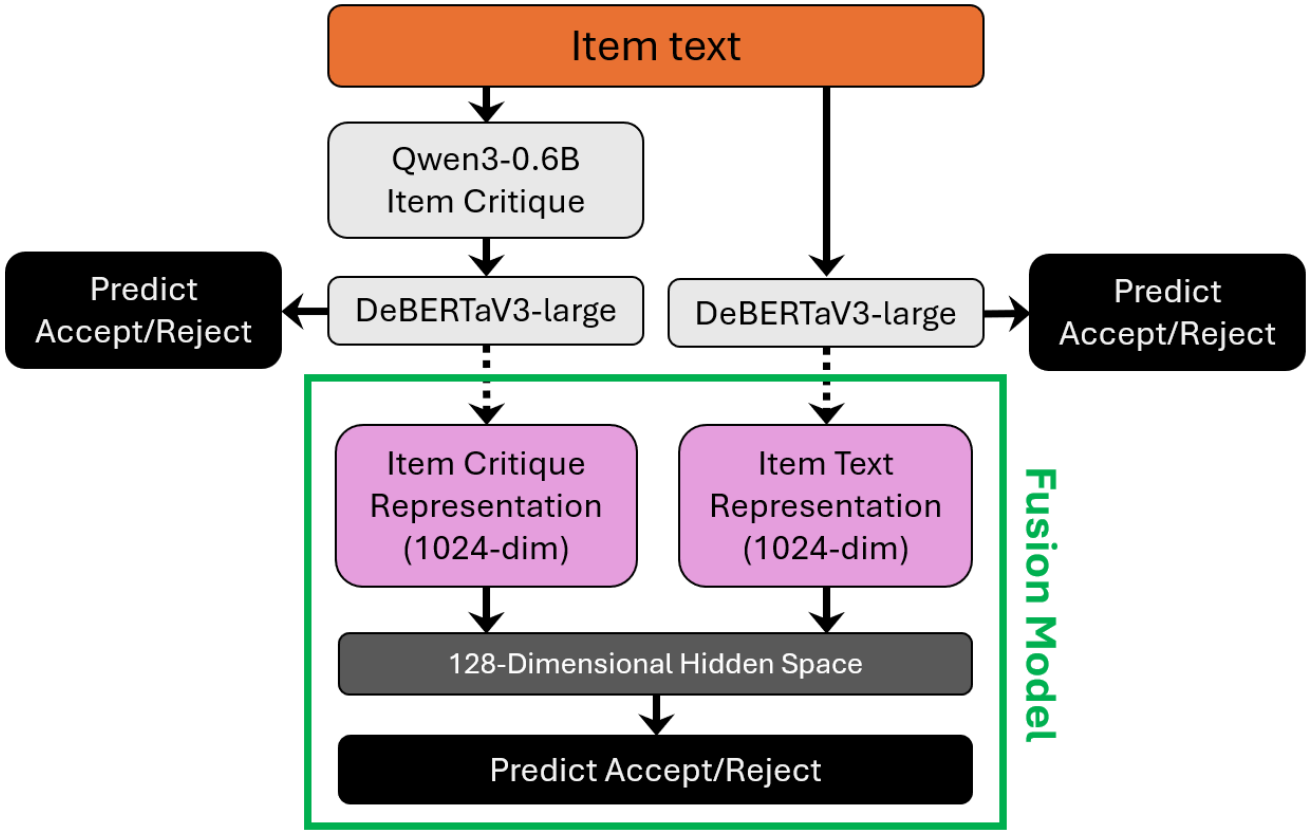
3.2.5 Stand-Alone Subject Models

Given that ELA and math are very distinct subjects, we included an alternative approach of building separate models for these subjects. All raw text-only, critique-only, and fusion models were retrained using this approach.

3.3 Evaluation

The area under the ROC curve (AUC) was used as a cutoff threshold-independent measure of the

Figure 1: Automated item evaluation model using raw text + LLM Critique fusion.



Note. Item text is processed directly through the first Deberta. Qwen is used to generate item critiques, then entered into a separate Deberta. Two DeBERTas are combined in the final fusion model.

model's overall discriminative ability. Classification thresholds were fixed to .50 by default. Using this threshold, we report accuracy (proportion of correctly classified items), precision (proportion of flagged items that are truly rejected), sensitivity (proportion of rejected items correctly flagged), specificity (proportion of viable items correctly passed), and F1 score (harmonic mean of precision and sensitivity). Similar to outlier detection, this paper treats rejected cases as positive cases for evaluation purposes.

3.3.1 Rejection Reason Classification

Item developers often left comments when moving an item to a permanent rejection status. Comment quality varied widely, ranging from absent or uninformative (e.g., "Moving to rejected") to detailed explanations. Multiple comments for the same item were concatenated and treated as a single record. Using a combination of Claude Sonnet 4.6 (Anthropic, 2025) and manual fixes, comments from rejected items were cleaned and classified into 10 rejection reason categories: content, psychometric, data review, bias, abandoned, incomplete, passage, scoring, non-content, and no data (see Appendix A for definitions). These labels were used to assess how well the best-performing model detected items across each rejection reason. The rejection reason data were only used for test sets. Only 19% of rejected test items had usable rejection reason data. Other items were classified in the "no data" category because the comments were missing (80%) or were ambiguous (1%).

3.4 Rejection Rate by Difficulty and Discrimination

We examined the Spearman correlations between the fusion model predictions and several item difficulty and discrimination metrics, including proportion correct, item-total correlation, and item response theory (IRT) parameters. All items were calibrated using the two-parameter logistic (2PL)

model (Birnbaum, 1968) for dichotomous items or the generalized partial credit model (GPCM; Muraki 1992) for polytomous items. For GPCM items, the item difficulty parameter was centered to represent average difficulty of the thresholds. Therefore, both item discrimination (a) and difficulty (b) parameters were comparable between 2PL and GPCM. The data spanned two subjects and multiple grades, but we standardized θ to have a mean of zero and standard deviation of one within each subject-grade combination, effectively removing the vertical scale for this analysis. Therefore, including proportion correct, item-total correlation, all item statistics are relative to the subject and grade. For example, a low proportion correct for an ELA grade 3 item is difficult for grade 3 students.

3.5 Sentiment Analysis of Critiques

To examine if sentiment can partly explain how the Qwen3 critiques improved prediction, we conducted a sentiment analysis. We computed sentiment using SiEBERT (Hartmann et al., 2023), a RoBERTa-large model fine-tuned for binary sentiment classification across 15 heterogeneous English text sources, chosen for its cross-domain generalization relative to models trained on a single text type. For each critique, we extracted the difference between the positive and negative class logits, yielding a continuous log-odds sentiment score in which higher values indicate more positive sentiment. We then computed Spearman correlations between critique sentiment and predicted rejection probability of the text-only model P_{text}, fusion model P_{fusion} and the difference $P\Delta = P_{fusion} - P_{text}$.

4 Results

All results are based on test data. Model performance is summarized in Table 1. Based on F1 score, the zero-shot approach was clearly the worst performer (F1 = .23). The fusion model consistently outperformed both the raw text-only and critique-only models (Accuracy = .75, F1 = .64, AUC = .80, precision = .63, sensitivity = .64, specificity = .81). In all cases, prediction for math was considerably more accurate than ELA. The stand-alone fusion models separated by subject (ELA and math) did not meaningfully change performance metrics. Note that a fusion model with non-weighted CEL had essentially the same performance and was not worth reporting separately (Accuracy = .75, F1 = .63, AUC = .80, precision = .62, sensitivity = .65, specificity = .79). However, the weighted model offers an interpretive advantage: because class weighting shifts the optimal decision boundary near 0.5, the classification threshold aligns with the intuitive probability midpoint. Subsequent analyses focus on evaluating the fusion model performance in detail.

4.1 Improving Sensitivity by Lowering the Cutoff Threshold

Lowering the cutoff threshold below .5 increases sensitivity to flag non-viable items at the cost of specificity (see Figure 2). For instance, a threshold of .25 raises the overall sensitivity to .90 while reducing specificity to .42 and F1 to .60. Math would have a sensitivity of .91 (specificity = .56, F1 = .70), while ELA would have a sensitivity of .88 (specificity = .31, F1 = .51; see Appendix E). A low detection threshold may be preferable in AIG, where producing items is far less costly than evaluating them.

4.2 Item Rejection Reasons

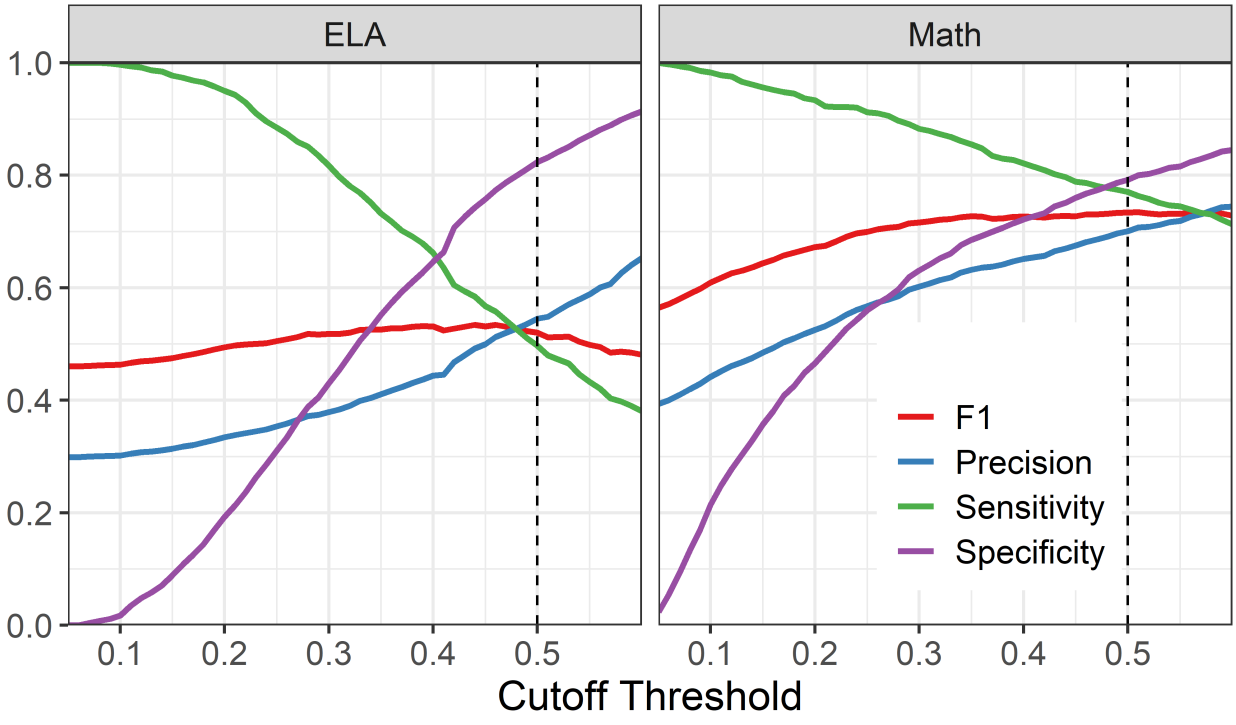
Predicted rejection probability by rejection reason was estimated on truly rejected items in the test data (see Figure 3). Items that were abandoned (sensitivity = .75) or had psychometric issues (sensitivity = .66) had one of the highest sensitivities for the fusion model, especially for math. Compared to the text-only model, incorporating item critiques in the fusion model improved detection across nearly all rejection categories. Low sensitivity (.16) to detect non-content rejections was expected as they typically reflect back-end data integrity issues unrelated to item text. Detection rate increased substantially for incomplete math items (Δ sensitivity = .31) and those with content issues (Δ sensitivity = .33). However, the fusion model struggled to detect concerns regarding bias, sensitivity, fairness, and accessibility (sensitivity = .27), most of which were ELA (91%).

Table 1: Classification Performance by Model and Subject

Model	p	AUC	Accuracy	Precision	Sensitivity	Specificity	F1
Zero-shot	.17	.50	.61	.35	.17	.83	.23
Raw text-only	.31	.76	.73	.61	.56	.82	.59
ELA	.21	.69	.72	.55	.39	.86	.46
Math	.42	.80	.74	.65	.71	.76	.68
Critique-only	.30	.72	.71	.59	.52	.82	.55
ELA	.20	.63	.69	.49	.32	.85	.39
Math	.41	.79	.74	.65	.68	.77	.67
Fusion (raw text + critique)	.34	.80	.75	.63	.64	.81	.64
ELA	.27	.72	.73	.54	.49	.83	.51
Math	.43	.86	.78	.70	.77	.79	.73
ELA Fusion (stand-alone)	.33	.69	.69	.47	.53	.75	.50
Math Fusion (stand-alone)	.39	.85	.79	.72	.73	.82	.73

Note. ELA = English language arts, p = proportion of items predicted as reject, AUC = area under the curve. Cutoff threshold for models were fixed to .5.

Figure 2: Cutoff threshold analysis for fusion model (raw text + critique)

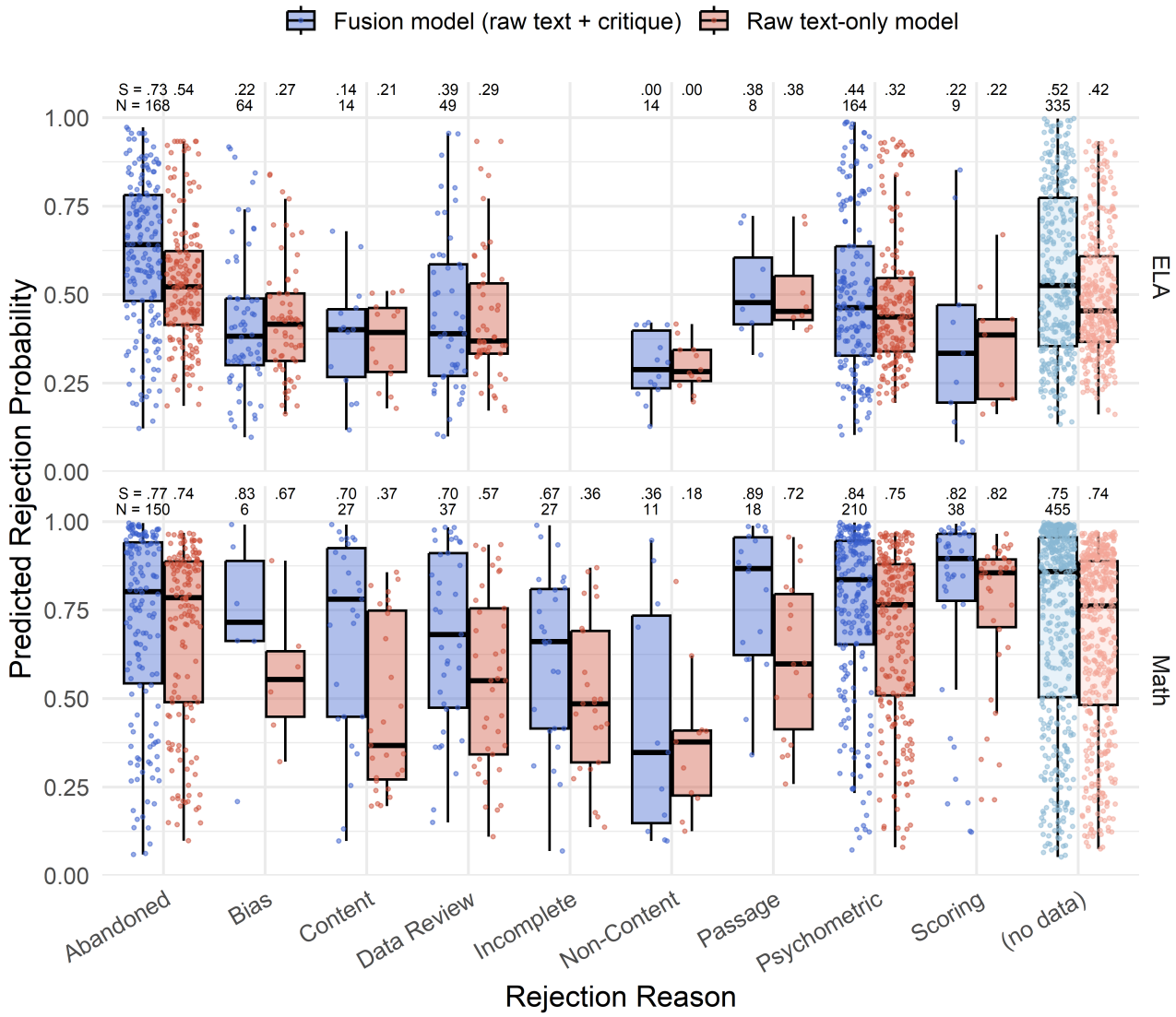


Note. ELA = English language arts. Lowering the cutoff threshold below .5 increases sensitivity to flag rejected items at the cost of specificity. A lower cutoff threshold may be preferable in contexts such as automated item generation where item generation has a low cost.

4.3 Item Difficulty and Discrimination

Spearman correlations between predicted rejection probability and several psychometric statistics was strongest with proportion correct ($r = -.30$), indicating that the model assigned higher rejection probabilities to more difficult items. Correlations with total item correlation ($r = -.06$), IRT discrimination ($r = -.07$), and the item intercept parameter ($-1.7ab$; $r = -.19$) were weaker. Overall, the model is sensitive to rejecting items that may

Figure 3: Boxplots of predicted rejection probability by rejected reason for items truly labeled as rejected.



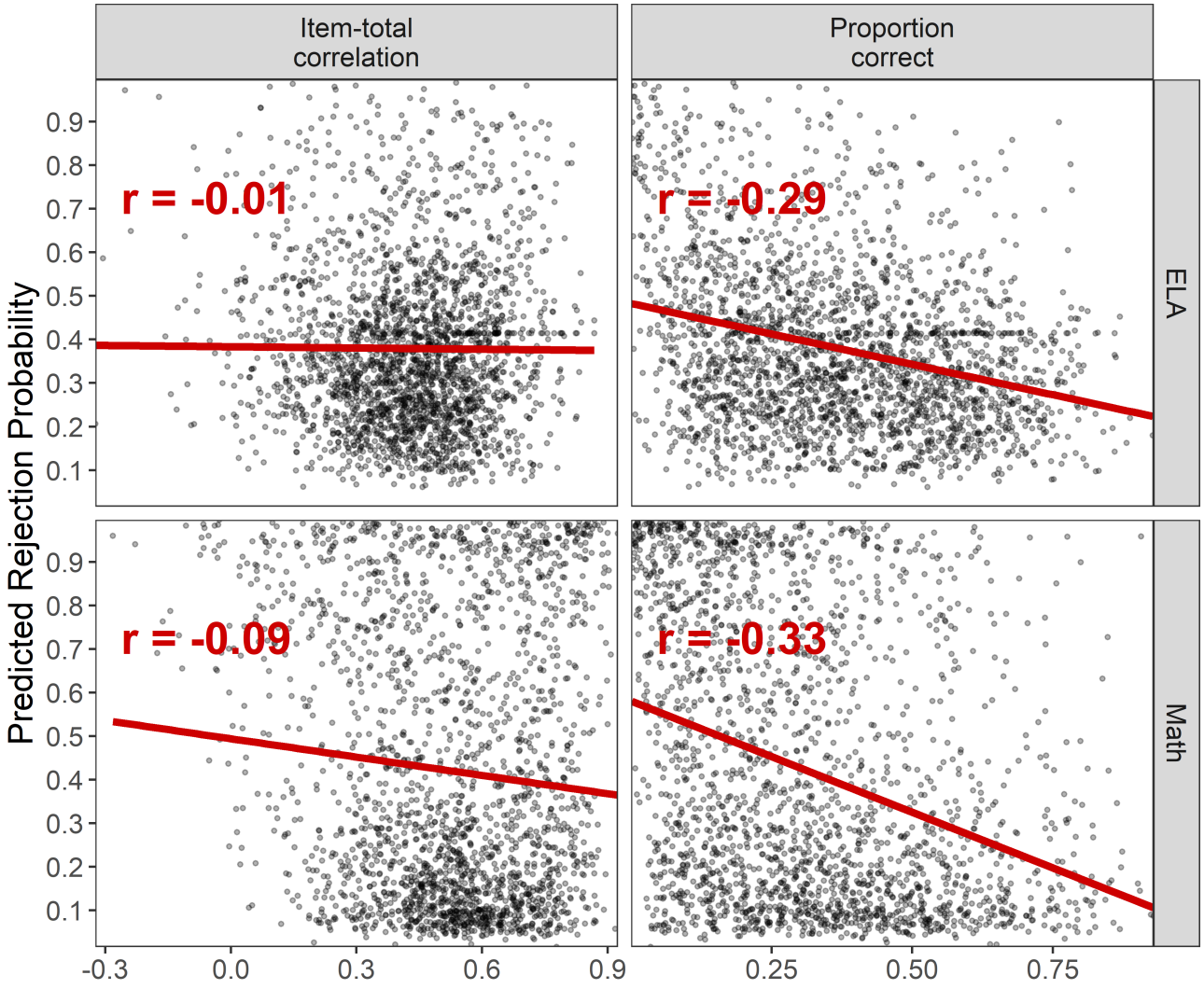
Note. ELA = English language arts. S=Sensitivity (using $\geq .5$ probability threshold), N=number of items, Scoring = issues with rubrics, answer keys, scoring logic, or scoring metadata, Abandoned = items without a suitable exam placement, Passage = rejection driven by the passage set rather than the item itself, Psychometric = psychometric issues like low discrimination and extreme difficulty, Incomplete = items that never reached field-testing due to stalling in development, Data review = questionable psychometric qualities requiring content expert review, including C-level differential item functioning, Content = content errors like standard misalignment, Bias = content accessibility, bias, fairness, or sensitivity concerns, Non-content = issues like missing, incompatible, or corrupt data that is not expected to be detectable by text-based models, (no data) = items rejected without comments or with unclear reasons. Only items truly labeled as rejected are shown. For full rejection reason details, see Appendix A.

be too difficult for the target population (see Figure 4). Note that 6.4% of items had missing item statistics because they were rejected prior to field test analyses. For these items, sensitivity was 66%, which was the same as the other items.

4.4 Critique Sentiment

Correlations between critique sentiment and P_{text} and P_{fusion} was significant for only math P_{fusion} ($r = -.05$, $p < .05$). Negative critique sentiment was associated with slightly increased $P\Delta$, across all subject and status label groups. The Spearman

Figure 4: Item difficulty and discrimination by fusion model-predicted rejection probability.



Note. r = Spearman correlations.

correlation between sentiment and $P\Delta$ was $-.05$ for ELA ($p < .01$) and $-.14$ for math ($p < .0001$). Disaggregating by status label, correlations were not significantly different from zero for ELA items ($p > .05$), and $-.11$ for accepted and $-.18$ for rejected math items ($p < .0001$). Therefore, the association was stronger for math than for ELA, and strongest among math items that had a true rejected status. Results were consistent with the interpretation that the critiques contributed information beyond the item text itself, where negative sentiment increased rejection rates. But overall, sentiment was a minor characteristic of how critiques were related to the model behavior.

5 Discussion

We began the study with the goal of building a comprehensive AIE model that could identify unwanted items across as many aspects of item quality as possible. We achieved this by training a transformer language model to classify items as accepted or rejected for operational use, using item status data from a large-scale standardized testing program. Data augmentation using item critiques generated by Qwen3 improved prediction accuracy across nearly all rejection reasons. The result was a single AIE model that is nearly comprehensive, much more convenient than most other AIE meth-

ods that target a single aspect of item quality. Our approach showed practical promise for large-scale assessment pipelines where both the evaluation and discard of items carry substantial cost.

We want to emphasize that AIE is useful for both human-written and AI-generated items, though the purpose and strategy of its use can vary. In AIG contexts, where item generation is inexpensive, a lower detection threshold such as .25 may be preferable. At the cost of specificity (.42) and F1 (.60), sensitivity of our fusion model rises to .90, catching most future rejections before they happen. This can save a substantial amount of time spent on content expert review. In more traditional item development contexts, the default threshold of .5 may serve as an effective early warning to revise the item. The predictions were considerably more accurate for math than ELA, which may suggest fundamental differences in how items are rejected between these subjects. Potentially, ELA item rejections were often related to their passages, which were excluded from the current study.

Among all item statistics examined, proportion correct (item difficulty) showed the strongest relationship with rejection. In spite of not directly using item statistics in the training, the model especially tended to flag items that were too difficult. This shows the value of item difficulty prediction research (AlKhuzaey et al., 2024; Benedetto et al., 2023), but the modest correlation we found ($r = -.30$) suggests that difficulty prediction alone is insufficient, as items are rejected for many reasons beyond difficulty.

Several aspects of this study are novel. First, the approach of fusing an encoder representation of raw item text with an encoder representation of a separate decoder-generated critique is relatively new, though a handful of studies have taken similar approaches (Henrichsen and Krebs, 2025; Hsieh et al., 2023; Scarlatos et al., 2025; Feng et al., 2025). This method leverages the decoder's capacity to draw on broad world knowledge when generating critiques. Decoders are not designed to output numerical estimates reliably, so we rely on the encoder for precise, task-specific prediction. Our approach could likely improve by using a larger decoder. The Qwen3-0.6B model we used is very small (i.e., 600 million parameters) compared to the most powerful proprietary multimodal models rumored to have trillions of parameters (Li, 2026). The caveat is that these proprietary models may require additional security measures unlike models that can be ran locally like Qwen. The field of AIE could therefore benefit from testing more powerful decoders combined with efficient, precisely fine-tuned encoders.

Another novel aspect of this study was the use of item developer comments. We are not aware of any other paper that has used item developer comments to help build a prediction model in the educational assessment context. The closest precedent may be Ma (2025), who applied DistilBERT and machine learning to examinee comments collected after test administration, building a model to identify comments most relevant for item review. It established that free-text commentary contains information for item quality decisions. We considered using item developer comments in the training data but decided against it, as only 19% of our data had rejection comments. Nevertheless, we could benefit from further exploration of how item developer comments can be leveraged for AIE.

Our model takes a nearly comprehensive approach to AIE, as the key outcome of judging item quality is whether it is acceptable for operational use. However, we cannot fully claim comprehensiveness for several reasons. For example, item quality partly depends on the other items already in the bank. We do not want two or more items that are too similar to each other, and detecting this would require a similarity analysis (Peng, 2020). Further, the distribution of item difficulty within a bank or test should typically be balanced (van der Linden and Pashley, 2000). Evaluating these require comparing the new items to the current item bank.

Finally, we acknowledge that some rejection reasons are inherently unpredictable from item content alone, including item exposure, rejection due to other items within the same passage set, data integrity issues, or dependence on future events not yet reflected in the data. Incorporating full item metadata including the passage text and scoring rubric into the prediction could partly resolve this, but some item quality issues will always depend on unpredictable future events.

One of the biggest concerns with our fusion model was its difficulty identifying items flagged for bias, sensitivity, fairness, or accessibility concerns,

possibly because these judgments require contextual and cultural knowledge that is difficult to encode from item text alone. This suggests that human review remains essential for detecting these issues in particular.

6 Conclusion

We introduced a novel approach to automated item evaluation (AIE) by training a transformer language model to predict item acceptance or rejection using item status data from a large-scale standardized testing program. Augmenting item text with LLM-generated critiques improved prediction accuracy across nearly all rejection reasons, demonstrating that combining encoder and decoder representations is a promising direction for AIE. Unlike prior work that has focused on predicting individual item properties such as difficulty, our approach targets the operational decision that ultimately determines an item's fate, offering a single, near-comprehensive quality metric applicable to both human-written and AI-generated items. As AIG continues to expand the volume of items requiring review, scalable AIE methods like the one presented here offer practical promise for reducing the cost and burden of item evaluation in large-scale assessment programs.

References

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.

AlKhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. (2023). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, pages 1–53.

AlKhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Amini, M., Ahmadi, B., Xiong, X., Zhang, Y., and Qiao, C. (2025). Prompting strategies for language model-based item generation in k-12 education: Bridging the gap between small and large language models.

Anthropic (2025). Claude Sonnet 4.6. `https://www.anthropic.com`. Large language model.

Benedetto, L., Cappelli, A., Turrin, R., and Cremonesi, P. (2020). R2de: a nlp approach to estimating irt parameters of newly generated questions. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 412–421.

Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., and Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R., editors, *Statistical theories of mental test scores*, pages 397–479. Addison-Wesley.

Circi, R., Hicks, J., and Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education*, Volume 8 - 2023.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1.

Falcão, F., Costa, P., and Pêgo, J. M. (2022). Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, 27(2):405–425.

Feng, W., Tran, P., Sireci, S., and Lan, A. (2025). Reasoning and sampling-augmented mcq difficulty prediction via LLMs. *arXiv preprint arXiv:2503.08551*.

Fu, Y., Jiao, H., Zhou, T., Zhang, N., Li, M., Xu, Q., Peters, S., and Lissitz, R. W. (2025). Text-

based approaches to item alignment to content standards in large-scale reading & writing tests.

Gorgun, G. and Bulut, O. (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.

Haladyna, T. M. and Downing, S. M. (1989). Taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1):37–50.

Han, S., Rijmen, F., Boykin, A. A., and Lottridge, S. (2025). Leveraging fine-tuned large language models in item parameter prediction. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 250–264.

Hartmann, J., Heitmann, M., Siebert, C., and Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

He, P., Gao, J., and Chen, W. (2021). Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Henrichsen, M. and Krebs, R. (2025). Two-stage reasoning-infused learning: Improving classification with LLM-generated reasoning. *arXiv preprint arXiv:2507.00214*.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. (2023). Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Hsu, F.-Y., Lee, H.-M., Chang, T.-H., and Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Kolesnikova, D., Fedyanin, K., Hofman, A. D., Brinkhuis, M. J., and Bolsinova, M. (2026). Estimating item difficulty with large language models as experts. *arXiv preprint arXiv:2605.18562*.

Li, B. (2026). Incompressible knowledge probes: Estimating black-box LLM parameter counts via factual capacity. *arXiv preprint arXiv:2604.24827*.

Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., and Lissitz, R. W. (2025). Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*, 85(6):1065–1090.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N. D., and Xu, M. (2025). Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.

Ma, Y. (2025). Analyzing examinee comments using DistilBERT and machine learning to ensure quality control in exam content. *arXiv preprint arXiv:2504.06465*.

Maeda, H. (2025). Field-testing multiple-choice questions with AI examinees: English grammar items. *Educational and Psychological Measurement*, 85(2):221–244.

Maeda, H. and Lu, Y. (2025). Finding words associated with DIF: Predicting differential item functioning using LLMs and explainable AI. *Journal of Educational Measurement*, 62(4):883–906.

Maeda, H. and Lu, Y. (2026). Multimodal test item parameter prediction from text, images, and metadata: Fusing together AI vision and language models. *Educational and Psychological Measurement*.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series*, 1992(1):i–30.

Oluoke, O. M., Gorbacheva, A. M., and Monday, O. A. (2026). Evaluating LLM-generated assessment items: A JBI-guided critical appraisal and checklist. *International Journal of Evaluation and Research in Education*. In press.

OpenAI (2024). GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Pelánek, R., Effenberger, T., Kukučka, A., et al. (2022). Towards design-loop adaptivity: identifying items for revision. *Journal of Educational Data Mining*, 14(3):1–25.

Peng, F. (2020). *Automatic enemy item detection using natural language processing*. PhD thesis, University of Illinois Chicago.

Prentzas, J. and Binopoulou, A. (2025). Explainable artificial intelligence approaches in primary education: A review. *Electronics*, 14(11).

Qwen Team (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Scarlatos, A., Fernandez, N., Ormerod, C., Lottridge, S., and Lan, A. (2025). Smart: Simulated students aligned with item response theory for question difficulty prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25082–25105.

Shin, J. and Gierl, M. J. (2024). Automated short-response scoring for automated item generation in science assessments. In *The Routledge International Handbook of Automated Essay Evaluation*, pages 504–534. Routledge.

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

van der Linden, W. J. and Pashley, P. J. (2000). *Item Selection and Ability Estimation in Adaptive Testing*, pages 1–25. Springer Netherlands, Dordrecht.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, Y., Gopalakrishnan, M., and Bergner, Y. (2025). Using generated rubrics to provide a window into item evaluation with multi-agent LLMs. In *International Conference on Artificial Intelligence in Education*, pages 203–217. Springer.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., and Poli, I. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Wauters, K., Desmet, P., and Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58(4):1183–1193.

Yaneva, V., Ha, L. A., Baldwin, P., and Mee, J. (2020). Predicting item survival for multiple choice questions in a high-stakes medical exam. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6812–6818, Marseille, France. European Language Resources Association.

Yu, M. C. and Burke, M. I. (2026). Automatic item generation, evaluation, and scale construction of non-cognitive measures with generative language models. In Thompson, I., Yankov, G. P., and Hernandez, I., editors, *Artificial Intelligence for I-O Psychologists: Research and Applications*. Oxford University Press.

Appendix A Rejection Reason Definitions

1. Content: content issue or error, including but not limited to (1) misalignment with the blueprint or standards, (2) unrealistic, inconsistent, or confusing scenarios, or (3) basic issues with the text or image such as grammar errors or missing information. A large portion of these items are rejected at a content review event just prior to field testing.

2. Psychometric: The following statistical criteria result in immediate rejection without manual review: proportion correct below .03, item-total correlation at or below .05, a distractor with a higher item-total correlation than the correct answer, poor inter-rater reliability on hand-scored items (exact agreement less than 75%, 65%, or 55% on items with 2, 3, or greater score points, respectively), or polytomous items where mean estimated θ does not increase monotonically across score levels. Items can also be rejected for less severe psychometric issues, though these require review and situational judgment: low or negative discrimination, too easy, too difficulty, insufficient sample size, overexposure, parameter drift, or differential item functioning.

3. Data review: These are items with questionable psychometric qualities that needed to be reviewed by content experts. Items flagged with C-level differential item functioning are always reviewed by content experts.

4. Bias: Content bias, sensitivity, fairness, or accessibility reasons

5. Abandoned: Items that could not find a suitable use in an exam. Many tend to be extra items in passage sets. Sometimes, the item itself has no content issues.

6. Incomplete: Items stuck in the item development process. Never reached field-testing.

7. Passage: The entire passage set (i.e., stimulus or testlet) was rejected. Sometimes, the item itself has no content issues.

8. Scoring: Any issue related to scoring rubric, answer key, scoring logic, or scoring meta data. This can be a combination of content, data integrity, and psychometric issues.

9. Non-content: Issues unrelated to the text or images of the item itself, such as missing metadata, corrupt or malformed data, format incompatibilities, import errors, or expired copyright permissions. Because these are primarily data integrity issues, text-based models are not expected to detect them.

10. (no data): Items rejected without comments or unclear reasons.

Appendix B Qwen3 Zero-Shot Classification Prompt

Below is an example python script for generating the Qwen3-0.6B zero-shot classification prompt.

```
SYSTEM_PROMPT = (
    "You are an expert assessment item reviewer. You will be given an assessment item.
    Respond with exactly one word: 'accept' or 'reject'. Respond 'accept' if the item is
    usable for assessment as-is. Respond 'reject' if the item has issues that go beyond a
    simple fix, such as: fundamental ambiguity, significant bias or sensitivity concerns,
    clear misalignment to the stated standard, or broken/unfixable formatting. Do not include
     any explanation, punctuation, or additional text."
)
def make_messages(row):
    return {
        "messages": [
            {"role": "system", "content": SYSTEM_PROMPT},
            {"role": "user", "content": f"This is a {row['subject']} subject, grade {row['
    grade']} item: \n{row['text']} \n{row['standards']} \nAny relevant images, figures, or
    tables have been excluded."
                }]}
```

Appendix C Qwen3 Item Critique Prompt

Below is an example python script for generating the Qwen3-0.6B item critique prompt.

```
SYSTEM_PROMPT = (
    "You are an expert assessment item reviewer with deep knowledge of psychometrics and item
     development best practices. You will be given an assessment item. Review the item for
    common quality issues such as: ambiguity, item difficulty, cultural or demographic bias
    and sensitivity, content alignment to standards, and flawed grammar or formatting. Write
    a concise 2 sentence summary: if the item has issues, name them plainly and briefly
    explain why they matter; if the item looks sound, say so and note its strongest quality.
    Do not use bullet points or headers. Be direct and specific - avoid vague praise or vague
     criticism."
)
def make_messages(row):
  return {
    "messages": [
            {"role": "system", "content": SYSTEM_PROMPT},
            {"role": "user", "content": f"This is a {row['subject']} subject, grade {row['
    grade']} item: \n{row['text']} \n{row['standards']}  \nAny relevant images and figures or
     tables have been excluded. \nWrite a concise 2 sentence summary about the item quality,
    including but not limited to its item difficulty, bias and sensitivity, and content
    alignment to standards."
            }]}
```

Appendix D Example Qwen3 Critiques

Table 2: Example Critiques Generated using Qwen3-0.6B.

Subject	Status	Grade	Critique	Sentiment	P_{fusion}	$P\Delta$
ELA	Accept	3	The item presents a grade 3 ELA question about identifying similarities between Bear and Max as cats, using plausible answers from the provided choices. It appears to be a standard reading comprehension task, with clear instructions and aligned content to educational standards. No ambiguities, biases, or content misalignment are evident.	6.8	.77	.44
	Accept	7	The item presents a clear conflict between the percentage-based data (Source 1) and the generalization (Source 2), creating ambiguity. While the standards are met, the item lacks contextual clarity on how the sources' findings interact, making it a flawed structure.	-6.0	.19	-.03
	Reject	4	The item contains ambiguity in the phrase "wide," leaving unclear alternatives for describing the wide chair. The correct response specifies "bumpy" instead, aligning with standards for precise language use. The item meets criteria for clarity and standard alignment.	6.6	.26	.02
	Reject	4	The item presents a question about the narrator's feelings toward being a seeing eye dog, but lacks clarity in the answer choices, leaving ambiguity about the intended response. It also includes an image/figure exclusion, which may impact the assessment's validity. No standard content alignment was identified.	-6.9	.80	.44
Math	Accept	6	The item lacks necessary visual elements (excluded images/tables) and fails to meet the standard "Reporting the number of observations," as the task requires entering a count instead of identifying the number of data points. It appears to align well with the specified standards but may require additional contextual clarity to ensure correctness.	5.2	.10	-.03
	Accept	7	The item presents a clear mathematical statement about how multiplication of coefficients relates to variables, making it suitable for grade 7 students. However, the phrasing "supports" or "does not support" lacks specificity, leaving ambiguity around the intended meaning. The question aligns with standard algebraic reasoning but requires further clarification to ensure accuracy. The item shows strong alignment with grade-level curriculum standards.	6.7	.80	.33
	Reject	7	The item contains ambiguity in the phrasing, making it unclear how the answer hinges on x's relationship to the context. The difficulty is moderate, aligning well with grade-level standards. The content is aligned to mathematical benchmarks without bias.	6.3	.99	.04
	Reject	8	The item contains a formatting error in the equation, specifically a missing fraction sign, which affects the interpretation of the algebraic manipulation. The answer choices provide conflicting explanations regarding the equation's solution, but the correct response is that the equation has exactly one valid solution, x = 2, which aligns with standard algebraic techniques. The item lacks clarity in presenting the process, potentially leading to confusion about the expected outcome.	-6.9	.92	.33

Note. Status = true status label, Sentiment = sentiment logit (positive = positive sentiment, negative = negative sentiment), P_{fusion} = fusion model probability of rejection, $P\Delta = P_{fusion}$ minus text-only model probability of rejection, ELA = English language arts. Two critiques from accepted and rejected ELA and math items were randomly selected, with the requirement that one had a near-zero $P\Delta$, and another had $P\Delta > .3$.

Appendix E Classification with .25 Cutoff

Table 3: Classification Performance by Model with .25 Cutoff Threshold

Model	p	AUC	Accuracy	Precision	Sensitivity	Specificity	F1
Raw text-only	.78	.76	.51	.41	.92	.30	.56
ELA	.84	.69	.42	.33	.94	.20	.49
Math	.71	.80	.61	.50	.91	.42	.64
Critique-only	.80	.73	.48	.39	.92	.26	.55
ELA	.92	.63	.35	.31	.96	.09	.47
Math	.67	.79	.62	.50	.88	.46	.64
Fusion (raw text + critique)	.69	.80	.58	.45	.90	.42	.60
ELA	.75	.72	.48	.35	.88	.31	.51
Math	.62	.86	.70	.57	.91	.56	.70
ELA Fusion (stand-alone)	.89	.70	.38	.32	.96	.13	.48
Math Fusion (stand-alone)	.59	.85	.71	.58	.88	.59	.70

Note. ELA = English language arts, p = proportion of items predicted as reject, AUC = area under the curve. Cutoff threshold for models were fixed to .25.