

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

Hotaka Maeda ^{1,2*}
Yikai Lu ³

¹Smarter Balanced, Santa Cruz, CA, USA

²University of California, Santa Cruz, Santa Cruz, CA, USA

³Department of Educational Psychology, University of Minnesota-Twin Cities, Minneapolis, MN, USA

Abstract

We built a near-comprehensive automated item evaluation model, predicting historical item acceptance or rejection from item text and Qwen3-generated critiques using 52,759 items from a large-scale testing program. Two DeBERTaV3 classifiers fused reached AUC .80 overall and .86 for math. Fairness-related rejections remained difficult, underscoring the need for human review.

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), and AIG review papers consistently urge for methods that can quickly filter poor quality items (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, such as item difficulty (AlKhuzayy et al., 2024), content alignment with the assessment blueprint (Fu et al., 2025), or differential item functioning (Maeda and Lu, 2025). To define this broader problem space, we use the term *automated item evaluation* (AIE; Yu and Burke, 2026), extending its prior use from a narrower scope. Here, 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.

The difficulty of AIE stems from the countless reasons an item can be unsuited for operational use. The Standards for Educational and Psychological Testing (AERA et al., 2014) articulate the requirements that define a valid, reliable, and fair item, and

Haladyna and Downing (1989) cataloged 43 item-writing rules covering stem construction, correct answer selection, and distractor quality. Training and deploying a separate prediction model for each rule is unrealistic, as model training requires substantial sample sizes and computational resources. Detecting item flaws separately may also be inefficient for rejecting undesirable items.

A comprehensive AIE approach is therefore desirable. One possibility is to use an existing item bank, 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 pass 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 as an AIE method. We train a transformer language model to classify items as accept or reject for operational use, using item status data from a large-scale standardized testing program, while augmenting the data using item critiques generated by an LLM. Model behavior is analyzed using historical item developer comments and sentiment analysis.

2 Related Works

2.1 Item Evaluation

Item evaluation is an important step during test development (AERA et al., 2014; Haladyna and Downing, 1989). Gorgun and Bulut (2025) summarized existing approaches into three categories: metric-based evaluations, post-hoc analysis, and human evaluations. Metric-based approaches typically compare items with a set of reference items,

*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

including NLP-based techniques such as word count, TF-IDF (Benedetto et al., 2020), cosine similarity for identifying enemy items, and measures of readability and text complexity (Yu and Burke, 2026; Amini et al., 2025). Pelánek et al. (2022) used an outlier detection mechanism to flag statistically deviant items. However, metric-based approaches are limited by their reliance on reference items, which can restrict the content domain of generated items, as valid items may be penalized simply because they *appear* different from reference items at the level of linguistic features (Gorgun and Bulut, 2025).

Post-hoc analysis approaches instead involve field testing to collect item response data from a representative sample of examinees. The Standards recommend meeting psychometric specifications related to item difficulty, item discrimination, and the absence of differential item functioning (AERA et al., 2014). However, this is costly and time-consuming, requiring a substantial sample size for reliable estimates. Finally, human evaluators can identify problems difficult to detect computationally, but human evaluation is subjective and time-consuming, and finding suitable experts is challenging.

In summary, metric-based approaches are automatic but limited in scope, while post-hoc analyses and human evaluations can be high-quality but resource-intensive. We argue in the next section that AIE can combine the advantages of all three while bypassing some of their limitations.

2.2 “Automated” Item Evaluation

The term “automated item evaluation” or similar 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 sparse usage and variation in scope, definitions of AIE 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. Among the three categories described by Gorgun and Bulut (2025), only the metric-based approach meets this criterion, as it eliminates reliance on field testing or human evaluators.

That said, results from post-hoc analyses or human evaluations could serve as labels for training AIE prediction models. Gorgun and Bulut (2025) trained an LLM to identify low-quality AIG items, but such an approach provides only a coarse-

grained evaluation, classifying items simply as *good* or *bad*. Pelánek et al. (2022) argues this may be a fundamentally ill-defined problem, as items warranting attention depend on context and are partly subjective. Items may also be problematic for many reasons, including linguistic or psychometric characteristics that human evaluators cannot always capture, and vice versa. Predicting statistical or psychometric properties from past post-hoc data has received more attention, including item difficulty (AlKhuyaey et al., 2024; Benedetto et al., 2023), item discrimination (e.g., Maeda and Lu, 2026; Han et al., 2025; Yaneva et al., 2020), and differential item functioning (Maeda and Lu, 2025). Although linguistic information was used as input, the outcomes were limited to psychometric characteristics rather than directly assessing linguistic or semantic dimensions of item quality.

LLM evaluators could be considered a fourth category. LLMs can mimic human evaluators but are automated, and unlike fine-tuning they require no labeled training data and can run zero-shot. Notably, LLM-based evaluation has enabled automated assessment of item-content alignment with assessment blueprints, an area that has received limited attention (Fu et al., 2025), suggesting LLM-based evaluation could shed light on aspects difficult to evaluate automatically.

The major limitation of the metric-based and psychometric-prediction approaches is that they often focus on one aspect of item quality at a time, even though different dimensions may be interrelated. For example, a psychometrically difficult mathematics item may also contain difficult words, a relationship that cannot be leveraged when these issues are evaluated separately. Relying solely on NLP or psychometric evaluations may also overlook problems identifiable only through human judgment. It is therefore preferable to construct a model that evaluates multiple dimensions simultaneously. Our comprehensive approach does so by using a dataset of rejected items with a broad range of documented rejection reasons. Rather than relying solely on human evaluators, psychometric analyses, or linguistic features, we fine-tune a Transformer language model to distinguish acceptable from unacceptable items while implicitly learning from multiple sources of item-quality evidence.

2.3 Transformer Language Models for Text Classification

The rise of Transformer language models increased demand for AIE research by making large-scale AIG feasible, while expanding the computational tools useful for AIE. Traditionally, text-based prediction in assessment 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 models have substantially improved accuracy (Li et al., 2025; Han et al., 2025; Maeda, 2025). Introduced by Vaswani et al. (2017), the Transformer replaced recurrent networks with parallel attention, enabling efficient training 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, with text tokenized into embeddings and passed through encoder layers to produce contextually enriched representations.

We use DeBERTaV3-large (He et al., 2021), which extends BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). DeBERTa disentangles content and positional information within its attention mechanism, representing each token by separate content and position vectors to capture relationships between words more effectively. The V3 variant adds replaced token detection, yielding a 304-million-parameter model whose natural language understanding even outperforms the newer ModernBERT (Warner et al., 2024).

2.4 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). In the assessment context, GPT-4o (OpenAI, 2024) was used to generate the reasoning steps required to reach each multiple choice option, which were then input into the Longformer encoder 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 difficulty (Kolesnikova et al., 2026).

We use a similar data augmentation technique with Qwen3 (Qwen Team, 2025). The Qwen3 follow the broader trend of capable small language models (SLMs) suited to resource-constrained set-

tings (Lu et al., 2025). A key capability distinguishing them from earlier SLMs is the integration of a *thinking mode* and a *non-thinking mode* within a single model, letting users toggle between slow, multi-step reasoning and fast response generation at inference time. Empirical results 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.

3 Methods

3.1 Item Data

The study included 52,759 items designed for English language arts (ELA) and mathematics state summative assessments for grades 3–11. Of these, 34% had 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, and spans a range of reasons including content and standards alignment, psychometrics, accessibility and sensitivity, and non-content issues such as data integrity problems (see Appendix A). Items that had not been rejected and had not yet been field tested were excluded, as their operational suitability remains unknown. Items rejected prior to field testing were retained, as their unsuitability is established (14% of rejected items). Items were field-tested among grade-matched students across multiple states in the United States.

To prepare for modeling, item text was concatenated with a \n separator between prompts and any available answer options. Selection-based (e.g., multiple choice, multiple select) and constructed response (e.g., short answer, extended response) items were included. Reading or listening passages were excluded due to their extensive length. Images, figures, tables, metadata, scoring rubrics, answer keys were also excluded. Data were partitioned randomly into approximately 80% training, 10% validation, and 10% test, with multiple items associated with the same stimulus always kept in the same group.

3.2 Prediction Approaches

Five prediction approaches were compared. All model training and inference used a single NVIDIA A10G Tensor Core 24GB graphics processor.

3.2.1 Zero-Shot Baseline

Qwen3-0.6B (Qwen Team, 2025) was prompted to classify each item as “accept” or “reject” directly, without task-specific fine-tuning (see Appendix B for the prompt). The subject, grade level, and aligned common core standards 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; rank = 32, 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 method that inserts trainable low-rank matrices into selected layers while keeping 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).

Accepted items outnumbered rejected items roughly two to one. This imbalance was addressed with class weights in the CEL, computed inversely proportional to class frequency so the minority class received greater influence on gradient updates. This preserves the full training sample while discouraging the model from defaulting to the majority class. We retained the state with the lowest validation CEL across 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, with subject, grade level, and aligned common core standards included in the prompt (see Appendix C for the prompt and Appendix D for example critiques). The resulting critique text was used as input to DeBERTaV3-large in place of the raw item text, with the same procedure and hyperparameters as the text-only model except LoRA rank = 50 and alpha = 100.

3.2.4 Text + Critique Fusion Model

The fusion model combines representations from both DeBERTaV3-large models above. The 1024-dimensional final-layer output of each was extracted and concatenated. The DeBERTa weights were frozen and only the fusion layers trained: 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

Because ELA and math are distinct subjects, we also re-trained separate raw text-only, critique-only, and fusion models for each.

3.3 Evaluation

The area under the ROC curve (AUC) measured overall discriminative ability independent of the cutoff threshold. With the classification threshold fixed at .50 by default, we report accuracy (proportion correctly classified), precision (proportion of flagged items truly rejected), sensitivity (proportion of rejected items correctly flagged), specificity (proportion of viable items correctly passed), and F1 (harmonic mean of precision and sensitivity). Rejected cases are treated as positive.

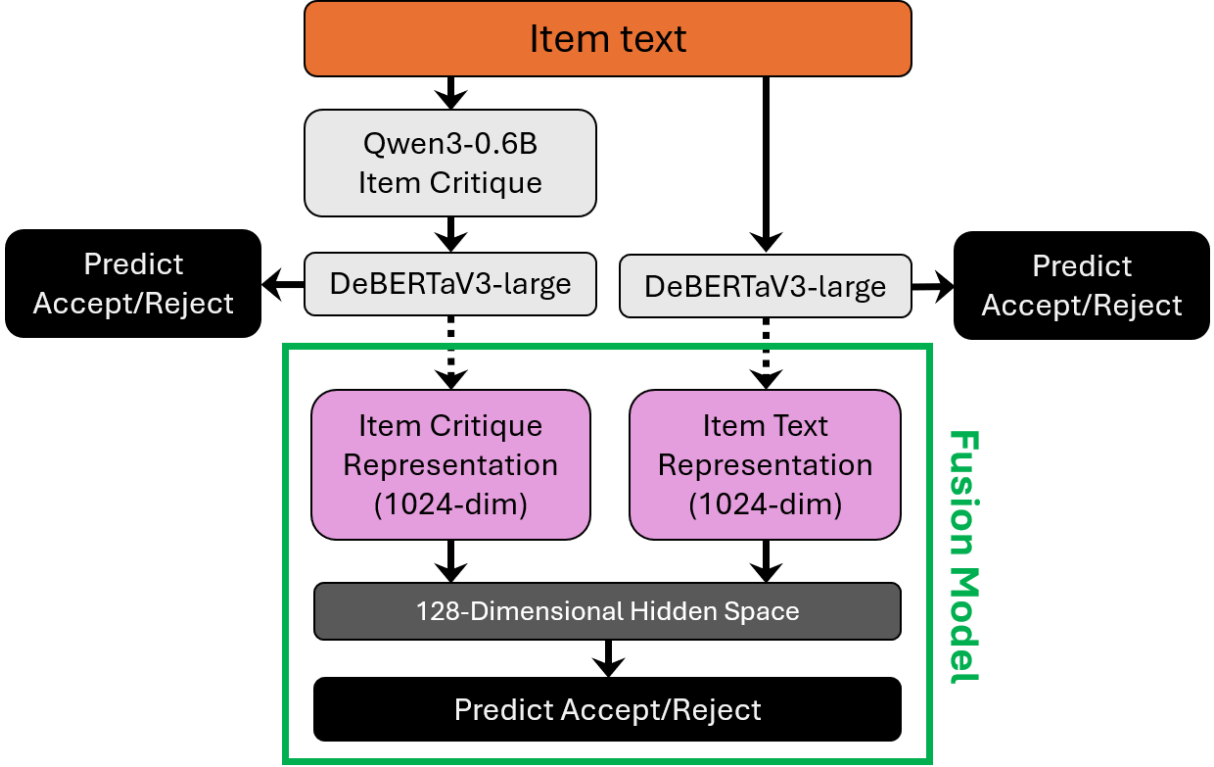
3.3.1 Rejection Reason Classification

Item developers often left comments when moving an item to permanent rejection, though quality varied from absent, uninformative (e.g., “Moving to rejected”), or detailed. Multiple comments for the same item were concatenated into a single record. Using Claude Sonnet 4.6 (Anthropic, 2025) with 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, used only for test sets, assessed how well the best-performing model detected items across each rejection reason. Only 19% of rejected test items had usable rejection reason data. The rest fell into “no data” because comments were missing (80%) or ambiguous (1%).

3.4 Rejection Rate by Difficulty and Discrimination

We examined Spearman correlations between fusion model predictions and several 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 difficulty parameter was centered to represent the average

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.

difficulty of the thresholds, so both discrimination (a) and difficulty (b) were comparable between 2PL and GPCM. Although the data spanned two subjects and multiple grades, we standardized θ to mean zero and standard deviation one within each subject-grade combination, removing the vertical scale for this analysis. All statistics, including proportion correct and item-total correlation, are therefore relative to subject and grade. For example, a low proportion correct for an ELA grade 3 item means it is difficult for grade 3 students. The 6.4% of items rejected prior to field testing were excluded.

3.5 Sentiment Analysis of Critiques

To examine whether sentiment partly explains how the Qwen3 critiques improved prediction, we computed sentiment using SiBERT (Hartmann et al., 2023). This is a RoBERTa-large model fine-tuned for binary sentiment classification across 15 heterogeneous English text sources. For each critique, we extracted the difference between the positive and negative class logits, where higher values indicate more positive sentiment. We computed Spearman correlations between critique sentiment and the pre-

dicted rejection probability of the text-only model P_{text} , the fusion model P_{fusion} , and their difference $P\Delta = P_{fusion} - P_{text}$.

4 Results

All results are based on test data. The zero-shot approach had the lowest F1 score ($F1 = .23$; see Table 1). The fusion model consistently outperformed both the raw text-only and critique-only models ($F1 = .64$). Prediction for math was considerably more accurate than ELA in all cases, and the stand-alone subject models did not meaningfully change performance. A fusion model with non-weighted CEL performed essentially the same ($F1 = .63$), but the weighted model offers an interpretive advantage: class weighting shifts the optimal decision boundary near 0.5, aligning the classification threshold with the intuitive probability midpoint. Subsequent analyses evaluate the fusion model.

4.1 Improving Sensitivity by Lowering the Cutoff Threshold

Lowering the cutoff below .5 increases sensitivity to flag non-viable items at the cost of specificity (see Figure 2). A threshold of .25 raises overall sen-

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.

sitivity to .90 while reducing specificity to .42 and F1 to .60. Math reaches sensitivity .91 (specificity = .56, F1 = .70) and ELA .88 (specificity = .31, F1 = .51; see [Appendix E](#)).

4.2 Item Rejection Reasons

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

4.3 Item Difficulty and Discrimination

The model assigned higher rejection probabilities to items with lower proportion correct ($r = -.30$). Correlations with total item correlation ($r = -.06$), IRT discrimination ($r = -.07$), and the intercept ($-1.7ab$; $r = -.19$) were weaker. Overall, the model is sensitive to rejecting items that may be difficult for the target population (see [Figure 4](#)).

4.4 Critique Sentiment

Correlations between critique sentiment and P_{text} and P_{fusion} were significant only for math P_{fusion} ($r = -.05$, $p < .05$). Negative critique sentiment

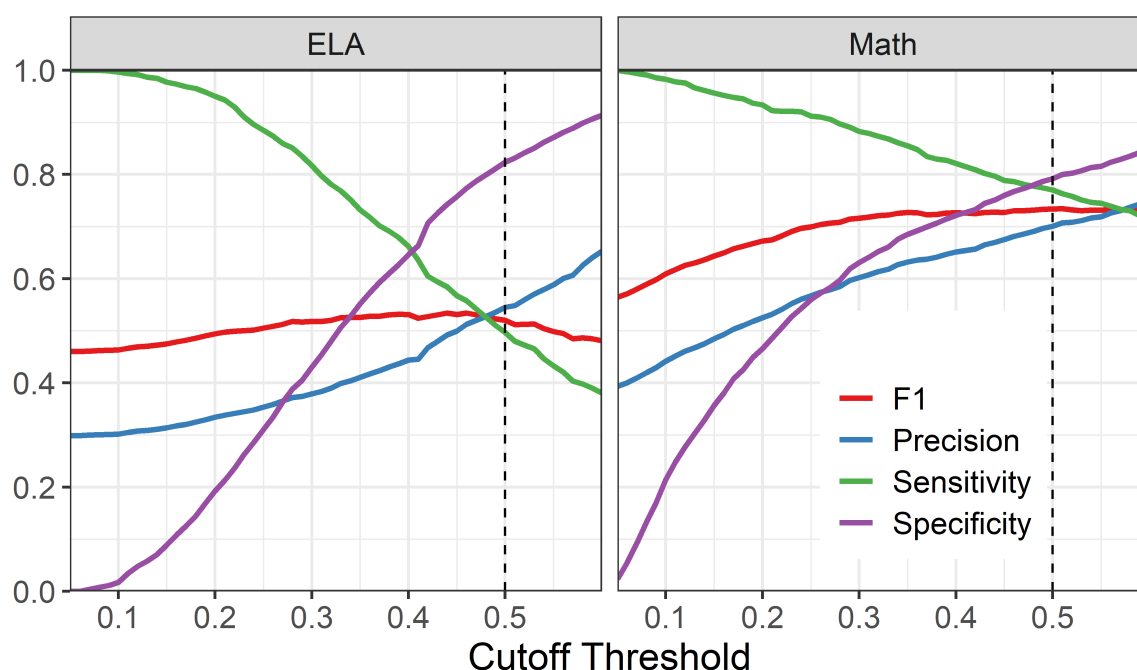
was associated with slightly increased $P\Delta$ across all subject and status groups. Correlation between sentiment and $P\Delta$ was $-.05$ for ELA ($p < .01$) and $-.14$ for math ($p < .0001$). Disaggregating by status, correlations were not significant for ELA ($p > .05$), and were $-.11$ for accepted and $-.18$ for rejected math items ($p < .0001$). The association was thus stronger for math than ELA, and strongest among truly rejected math items. Negative sentiment increasing rejection rates was expected, but sentiment was overall a minor characteristic of how critiques related to model behavior.

5 Discussion

We set out to build 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, near-comprehensive AIE model, more convenient than most other AIE methods that target a single aspect of item quality, with practical promise for large-scale assessment pipelines where both the evaluation and discard of items carry substantial cost.

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

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.

fusion model rises to .90, catching most future rejections before they happen and saving substantial content expert review time. In more traditional item development contexts, the default threshold of .5 may serve as an effective early warning to revise the item. Predictions were considerably more accurate for math than ELA, which may reflect fundamental differences in how items are rejected between subjects. In particular, ELA rejections were often related to their passages, which were excluded from the current study.

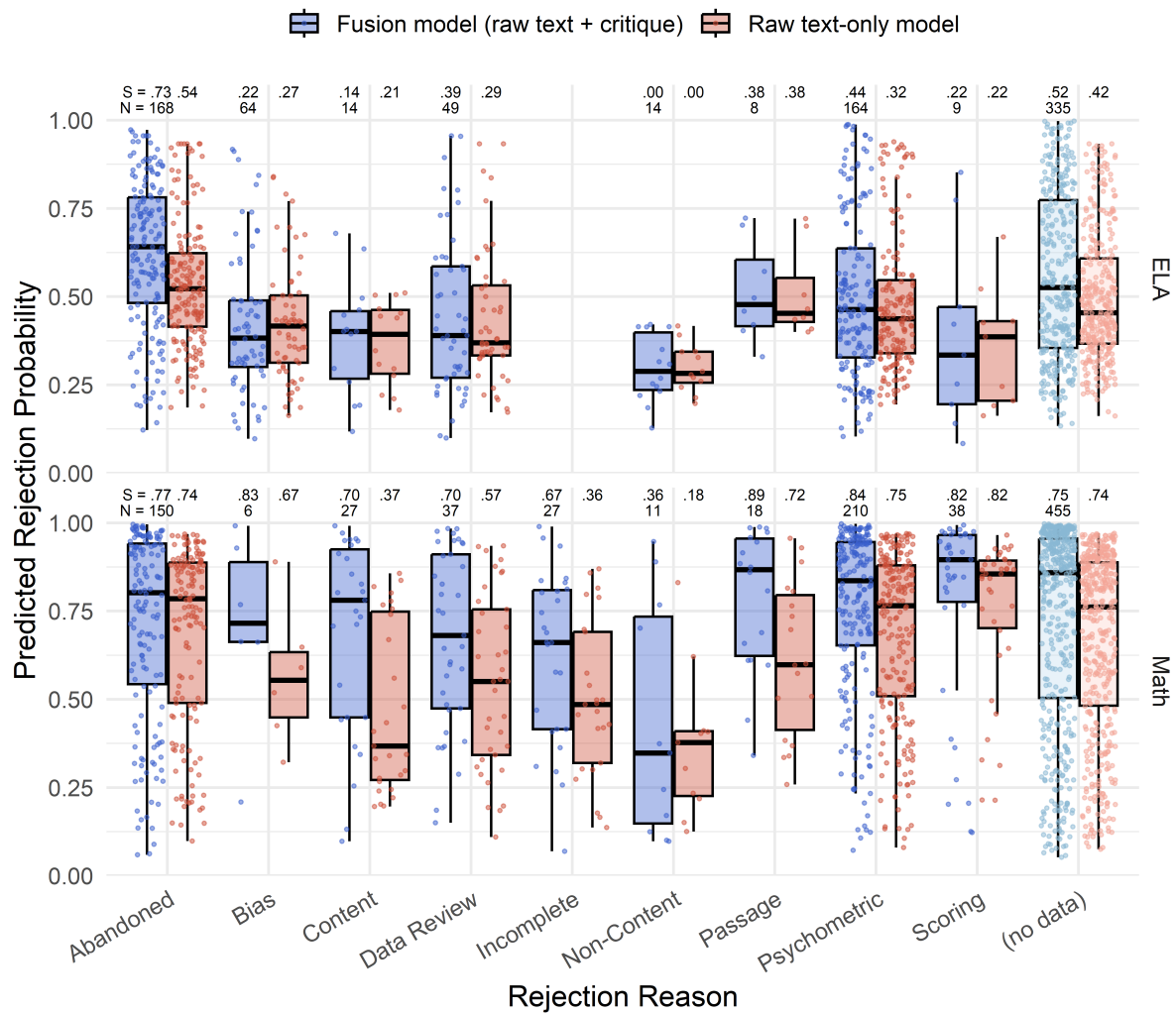
Associations with item statistics showed that the model rejected items that were too difficult. This shows the value of item difficulty prediction research (AlKhuzayy et al., 2024; Benedetto et al., 2023), but the modest correlation ($r = -.30$) suggests difficulty prediction alone is insufficient. Rejection rates were nearly unrelated to item discrimination, showing that items were rejected for many reasons beyond difficulty and discrimination.

Several aspects of this study are novel. First, 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, while relying on the encoder for the precise, task-specific prediction that decoders are not designed to produce. The approach could likely improve with a larger decoder, as our Qwen3-0.6B model is very small (600 million parameters) compared to proprietary multimodal models rumored to have trillions of parameters (Li, 2026), though such models may require additional security measures unlike locally run models like Qwen. Second, we are not aware of any other paper that has used item developer comments to help build a prediction model in educational assessment. The closest precedent may be (Ma, 2025), who applied DistilBERT and machine learning to examinee comments to identify those most relevant for item review, establishing that free-text commentary contains information for item quality decisions. We considered using item developer comments in training but decided against it, as only 19% of our data had rejection comments.

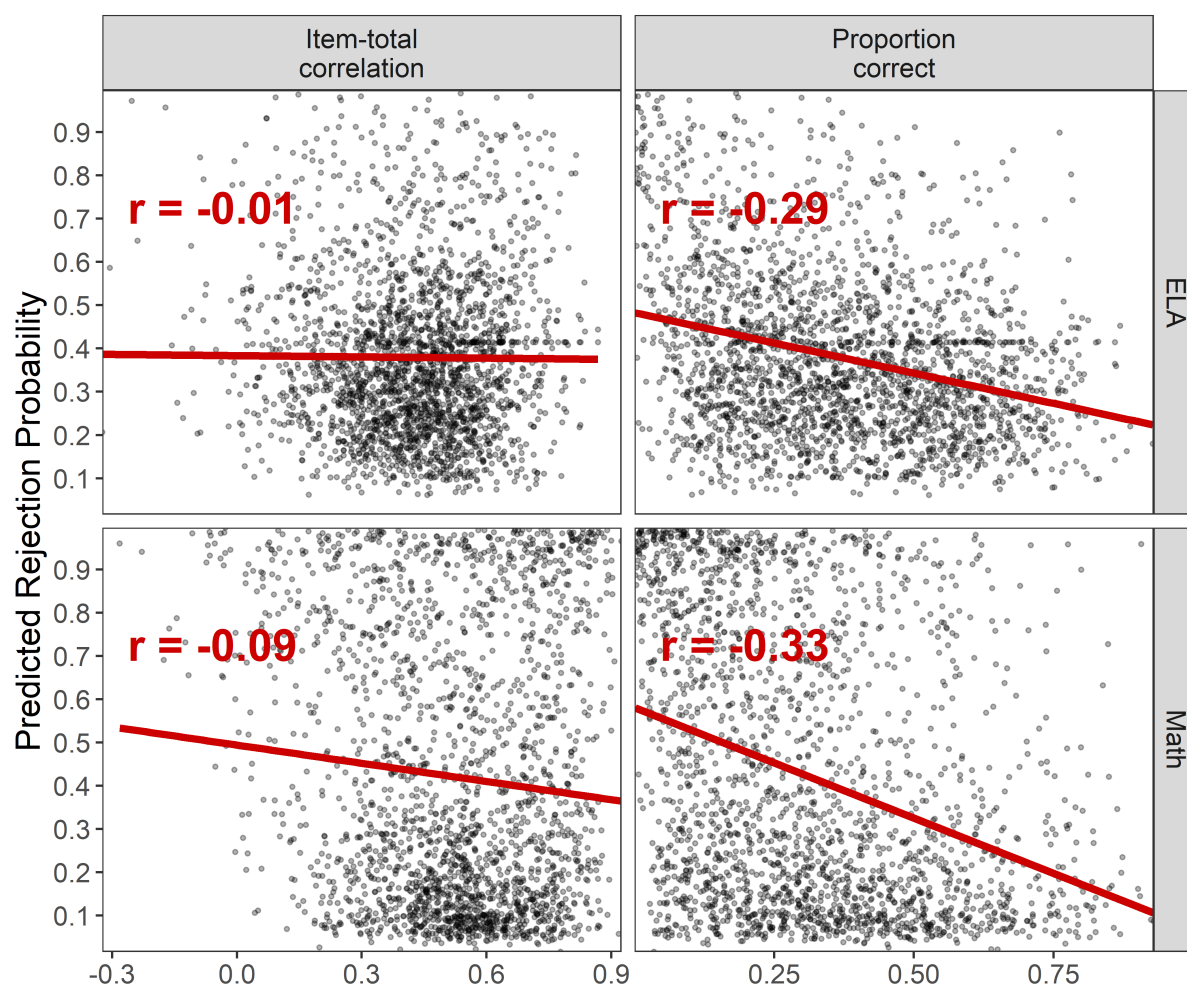
Our model takes a nearly comprehensive approach to AIE, as the key outcome of judging item quality is whether an item is acceptable for operational use. However, we cannot fully claim comprehensiveness. Item quality partly depends on the other items already in the bank. We do not

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](#).

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



Note. r = Spearman correlations.

want items that are too similar to each other, which would require a similarity analysis (Peng, 2020), and the distribution of item difficulty within a bank should typically be balanced (van der Linden and Pashley, 2000). Both require comparing new items to the current bank. Some rejection reasons are also 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. Incorporating full item metadata such as passage text and scoring rubric could partly resolve this, but some issues will always depend on unpredictable future events. Finally, our fusion model struggled to identify items flagged for bias, sensitivity, fairness, or accessibility concerns, possibly because these judgments require contextual and cultural knowledge difficult to encode from item text alone. This suggests human review remains essential for detecting these issues in particular.

Unlike prior work 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.
- Samah AlKhuzaei, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.
- Mohammad Amini, Babak Ahmadi, Xiaomeng Xiong, Yilin Zhang, and Christopher Qiao. 2025. [Prompting strategies for language model-based item generation in k-12 education: Bridging the gap between small and large language models](#). Preprint, arXiv:2508.20217.
- Anthropic. 2025. Claude Sonnet 4.6. <https://www.anthropic.com>. Large language model.
- Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 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.
- Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giusani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.
- A. Birnbaum. 1968. Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord and M. R. Novick, editors, *Statistical theories of mental test scores*, pages 397–479. Addison-Wesley.
- Ruhan Circi, Juanita Hicks, and Emmanuel Sikali. 2023. [Automatic item generation: foundations and machine learning-based approaches for assessments](#). *Frontiers in Education*, Volume 8 - 2023.
- Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 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.
- Filipe Falcão, Patrício Costa, and José M Pêgo. 2022. Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, 27(2):405–425.
- Wanyong Feng, Peter Tran, Stephen Sireci, and Andrew Lan. 2025. [Reasoning and sampling-augmented mcq difficulty prediction via LLMs](#). *arXiv preprint arXiv:2503.08551*.
- Yanbin Fu, Hong Jiao, Tianyi Zhou, Nan Zhang, Ming Li, Qingshu Xu, Sydney Peters, and Robert W. Lisitz. 2025. [Text-based approaches to item alignment to content standards in large-scale reading & writing tests](#).
- Guher Gorgun and Okan Bulut. 2025. [Instruction-tuned large-language models for quality control in automatic item generation: A feasibility study](#). *Educational Measurement: Issues and Practice*, 44(1):96–107.
- Thomas M. Haladyna and Steven M. Downing. 1989. [Taxonomy of multiple-choice item-writing rules](#). *Applied Measurement in Education*, 2(1):37–50.
- Suhwa Han, Frank Rijmen, Allison Ames Boykin, and Susan Lottridge. 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.
- Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. [More than a feeling: Accuracy and application of sentiment analysis](#). *International Journal of Research in Marketing*, 40(1):75–87.

- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *arXiv preprint arXiv:2111.09543*.
- Mads Henrichsen and Rasmus Krebs. 2025. [Two-stage reasoning-infused learning: Improving classification with LLM-generated reasoning](#). *arXiv preprint arXiv:2507.00214*.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. [Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes](#). *arXiv preprint arXiv:2305.02301*.
- Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Diana Kolesnikova, Kirill Fedyanin, Abe D Hofman, Matthieu JS Brinkhuis, and Maria Bolsinova. 2026. Estimating item difficulty with large language models as experts. *arXiv preprint arXiv:2605.18562*.
- Bojie Li. 2026. [Incompressible knowledge probes: Estimating black-box LLM parameter counts via factual capacity](#). *arXiv preprint arXiv:2604.24827*.
- Ming Li, Hong Jiao, Tianyi Zhou, Nan Zhang, Sydney Peters, and Robert W Lissitz. 2025. Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*, 85(6):1065–1090.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. 2025. [Small language models: Survey, measurements, and insights](#). *arXiv preprint arXiv:2409.15790*.
- Ye Ma. 2025. [Analyzing examinee comments using DistilBERT and machine learning to ensure quality control in exam content](#). *arXiv preprint arXiv:2504.06465*.
- Hotaka Maeda. 2025. [Field-testing multiple-choice questions with AI examinees: English grammar items](#). *Educational and Psychological Measurement*, 85(2):221–244.
- Hotaka Maeda and Yikai Lu. 2025. [Finding words associated with DIF: Predicting differential item functioning using LLMs and explainable AI](#). *Journal of Educational Measurement*, 62(4):883–906.
- Hotaka Maeda and Yikai Lu. 2026. [Multimodal test item parameter prediction from text, images, and metadata: Fusing together AI vision and language models](#). *Educational and Psychological Measurement*.
- Eiji Muraki. 1992. A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series*, 1992(1):i–30.
- OpenAI. 2024. [GPT-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- Radek Pelánek, Tomáš Effenberger, Adam Kukučka, et al. 2022. Towards design-loop adaptivity: identifying items for revision. *Journal of Educational Data Mining*, 14(3):1–25.
- Fang Peng. 2020. [Automatic enemy item detection using natural language processing](#). Ph.D. thesis, University of Illinois Chicago.
- Jim Prentzas and Ariadni Binopoulou. 2025. [Explainable artificial intelligence approaches in primary education: A review](#). *Electronics*, 14(11).
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 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.
- Jinnie Shin and Mark J Gierl. 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.
- Bin Tan, Nour Armoush, Elisabetta Mazzullo, Okan Bulut, and Mark Gierl. 2025. A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.
- Wim J. van der Linden and Peter J. Pashley. 2000. [Item Selection and Ability Estimation in Adaptive Testing](#), pages 1–25. Springer Netherlands, Dordrecht.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *arXiv preprint arXiv:1706.03762*.
- Yu Wang, Madhumitha Gopalakrishnan, and Yoav Bergner. 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.

- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *arXiv preprint arXiv:2412.13663*.
- Kelly Wauters, Piet Desmet, and Wim Van Den Noortgate. 2012. Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58(4):1183–1193.
- Victoria Yaneva, Le An Ha, Peter Baldwin, and Janet Mee. 2020. [Predicting item survival for multiple choice questions in a high-stakes medical exam](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6812–6818, Marseille, France. European Language Resources Association.
- Martin C Yu and Maura I Burke. 2026. [Automatic item generation, evaluation, and scale construction of non-cognitive measures with generative language models](#). In Isaac Thompson, Georgi P Yankov, and Ivan Hernandez, 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: low proportion correct, low item-total correlation, a distractor with a higher item-total correlation than the correct answer, poor inter-rater reliability on hand-scored items, 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 coherent analysis of Samantha’s emotional state, aligning well with grade 3 standards. It includes specific details about her actions and observations, making it suitable for assessing character responses to a narrative. No major quality issues are identified, as the content directly supports the intended learning outcomes.	6.8	.26	-0.03
ELA	Accept	3	The item contains a grammatical issue with repeated use of "sentence" and ambiguous phrasing, making clarity challenging. It aligns well with standards for grade 3 ELA, particularly focusing on analyzing stories and identifying central messages through textual evidence. The content is appropriate and includes elements like images and tables as specified.	6.8	.66	0.34
ELA	Reject	8	The item contains repetition (e.g., "very tangled" vs. "tangled") and lacks variation in language precision, making it a flawed grammar issue. It aligns with standard reading comprehension tasks but fails to address ambiguity in the task instructions.	-7.5	.73	0.39
ELA	Reject	11	The item presents a clear connection between Eisenhower’s plan for America’s development and his focus on global responsibility, aligning with the standard of analyzing how ideas interact over time. It correctly identifies key points such as the paragraph’s purpose to emphasize mutual trust and respect, making it a sound assessment item.	6.8	.37	0.02
Math	Accept	5	The item contains ambiguity in the number "about 90," which may leave students unsure if it refers to the total number of people or a specific group. It also shows potential difficulty in the answer key, requiring clarification on what data is needed to estimate waiting time. The question aligns with grade-level standards for mathematical estimation skills.	-4.9	.86	0.36
Math	Accept	6	The item lacks necessary visual elements (no image/table) and does not provide the required data (number of neighbors), making it ambiguous. It aligns with standards like reporting observations but fails to include supporting context.	-7.5	.14	-0.05
Math	Reject	6	The item presents a clear unit rate question about distance traveled, aligning with grade-level standards. It includes relevant data (3000 feet in one hour) without ambiguity, and the context is neutral, making it suitable for a math curriculum. No significant biases or flaws were identified.	6.8	.54	-0.02
Math	Reject	11	The item presents a valid mathematical scenario where the slopes of AE and BC are consistent with the properties of a right triangle, and all stated facts align with the given conditions. It does not contain ambiguities, biases, or content misalignment with educational standards. The highest quality is the clear application of slope relationships in a geometric context.	6.8	.70	0.44

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.