Leveraging Fine-tuned Large Language Models in Item Parameter Prediction

Suhwa Han[1], Frank Rijmen[1], Allison Ames Boykin[2*], Susan Lottridge[1]

[1]Cambium Assessment, [2]National Board of Medical Examiners

Correspondence: suhwa.han@cambiumassessment.com

Abstract

Accurate prediction of item parameters using item characteristics has been a long-standing objective in educational measurement, and recent advances in natural language processing (NLP) and large language models (LLMs) have opened new possibilities for modeling item parameters directly from item text. In this study, we introduce novel fine-tuning approaches that leverage item text as well as structured item attribute variables for enhanced prediction. For benchmarking, we compare suggested approaches with a traditional tree-based machine learning model that uses item attributes as primary inputs. The proposed methods are evaluated on a dataset of over 1,000 operational English Language Arts (ELA) items, with both dichotomous and polytomous scoring. Our work offers a unique opportunity to evaluate the prediction of item difficulty for polytomous items as well as item discrimination—areas that have received limited attention in prior research.

1 Introduction

Item parameter prediction in educational measurement refers to the modeling of item response theory (IRT) model parameters such as difficulty by using item-level features inherent in the items (AlKhuzaey et al., 2024). Accurately and reliably predicted item parameters offer multiple benefits. First, it can reduce the heavy reliance on field testing in evaluating new items, which is costly and increases the risk of security breaches due to item exposure (Ulitzsch et al., 2025). If non-functional items, such as those that are too easy or not discriminative, can be identified through predictive modeling, test developers can save resources in vetting new items. In addition, item parameter prediction has broader implications beyond large-scale assessments. If the methodology is sufficiently validated, it can be applied to classroom settings, where educators evaluate how their own items align with summative scales and make data-informed adjustments to instructions.

Given its potential to support a wide range of assessment activities, item parameter prediction has been a long-standing objective in the field (Fischer, 1995; Ferrara et al., 2022). In particular, recent advances in natural language processing (NLP) techniques has enabled researchers to leverage textual information in items in predicting item parameters (AlKhuzaey et al., 2024; Benedetto et al., 2023). Researchers have utilized language models to extract surface-level linguistic features and/or to derive embeddings to capture deeper semantic meanings, which are then used as features in statistical models or machine-learning (ML) algorithms (Xue et al., 2020; Yaneva et al., 2019, 2023). More recently, fine-tuning large language models (LLMs) on item texts has shown improved predictive performance compared to feature-based approaches (Benedetto et al., 2021; Yaneva et al., 2024; Zu and Choi, 2023). Given these recent findings and ever-improving capabilities of LLMs, further investigation into the use of fine-tuned LLMs for the item parameter modeling appears warranted and timely.

1.1 Study Purpose and Contributions

The purpose of the current study is to examine the performance of fine-tuned LLMs and explore how they can be more effectively leveraged for the item parameter prediction. To this end, the study addresses several research questions that have received limited attention in the existing literature. First, it investigates ways to integrate textual information from items with additional item attributes—such as content classification variables—within a fine-tuned LLM architecture. The study also compares the performance of fine-tuned LLMs with traditional machine learning algorithms to evalu-

*Work conducted while the author was at Cambium Assessment.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 250–264

ate their relative performance. Third, the study suggests methodologies for applying LLMs in the prediction of parameters for polytomously scored items, an area that has been underexplored. Finally, the study investigates the capacity of fine-tuned LLMs to predict discrimination parameters, which has historically received less attention in item parameter modeling.

2 Prior Research on IRT Parameter Prediction Using Fine-tuned LLMs

One of the key advantages of using LLMs is that they can be further fine-tuned for a specific task on top of its general linguistic capabilities obtained from pre-training. Through fine-tuning, the model parameters are optimized for a given task, enabling improved performance on downstream applications. Due to this flexibility and ability to directly model textual input, fine-tuned LLMs have been increasingly used to predict IRT parameters in the context of educational assessments.

Benedetto et al. (2021) demonstrated that a fine-tuned BERT model (Devlin et al., 2019) could effectively estimate difficulty parameters of Rasch model (Rasch, 1993) using items from e-learning platforms. They found that the fine-tuning approach reduced estimation error by 6.5% compared to traditional feature-based ML approaches using TF-IDF and embeddings. Zu and Choi (2023) also examined performance of fine-tuned RoBERTa model (Liu et al., 2019) in predicting item difficulty parameters of autogenerated multiple-choice items for English-language proficiency tests. By first fine-tuning RoBERTa on a key classification task subsequently adapting it for difficulty prediction, they achieved stronger correlations—$r = .733$ for listening and $r = .684$ for reading—compared to traditional methods based on hand-crafted features and embeddings.

Building on this line of work, Gombert et al. (2024) explored fine-tuning various transformer-based models to jointly predict both item difficulty and response time for multiplice-choice items in a medical licensure exam. They introduced architectural enhancements to LLMs by incorporating scalar mixing and a custom regression head. While their approach ranked first in a share task competition, their predictive power was relatively modest, yielding a maximum correlation of .27. Using a different dataset—math proficiency test data set for adults, Feng et al. (2024) found that fine-tuned RoBERTa achieved the best prediction, outperforming linear regression and zero-shot prompting approaches in terms of minimizing mean squared error, while explaining approximately 43% of the variance in the difficulty.

3 Methods

While several studies have successfully fine-tuned LLMs for item parameter prediction, to the best of the authors' knowledge, none have explored the integration of item attribute variables—such as content-wise classifications—directly within the LLM fine-tuning process. Given that most operationally maintained items are accompanied by such metadata, leveraging these additional features may enhance the predictive performance of LLMs.

In addition, the dataset used in this study is notable for its diversity, encompassing a range of item types that are currently operationally used in a large scale assessment. As such, evaluating the performance of the proposed methods on this dataset can provide insights that are both methodologically novel and practically relevant.

3.1 Dataset

The dataset used in this study consists of 1,119 items to assess English Language Art (ELA) proficiency for Grade 6 students. These selected items were drawn from the operational pool for the 2024-2025 Smarter Balanced assessment administration. The authors gratefully acknowledge the collaboration and support of Smarter Balanced in providing access to this high quality dataset.

These items span seven distinct item types, including five machine-scorable types (EBSR, HT, MC, MI, and MS) and two constructed response types (SA, WER) (see Appendix A for the description of the item types). In this dataset, the machine-scorable items were scored dichotomously, and constructed response items were all scored polytomously. The items were field-tested across 8 years (2014, 2015, 2016, 2017, 2018, 2019, 2020 and 2022), and include 935 summative items and 184 interim ones. The actual counts of the item types across the field-tested years can be found in Appendix B.

Two Sources of Information: Texts and Item Attributes. Each item in the dataset was associated with two types of texts: a stimulus text and as item text. Since these items were ELA items, stimulus texts typically consisted of a reading passage de-

signed to provide necessary information needed to answer questions. Item texts contained the actual question or prompt. To optimize input construction for the modeling, we concatenated the item text followed by the stimulus text. This ordering was to ensure inclusion of the item prompt within limited sequence length in the LLM modeling process. For polytomously scored constructed-response items, an additional piece of textual information was incorporated: rubric texts. The rubric texts were needed to provide unique information to model multiple difficulty parameters for the polytomous items. To ensure this critical information was retained in the modeling, rubric texts were prepended to the item text, followed by the stimulus text.[1]

In addition to textual data, this study extracted a set of item attribute variables to evaluate their contribution to the prediction. In total, 152 attribute variables were compiled: 59 content-based specification variables and 93 hand-crafted linguistic features extracted from both item and stimulus texts based on Baldwin et al. (2021) (see Appendix C for examples of item attribute variables used in this study.)

Target Variable: Banked IRT parameters. The target variables in this study were IRT parameters—both item difficulty and discrimination parameters—from the operational Smarter Balanced item bank. In the bank, the dichotomous items were calibrated using two-parameter logistic (2PL) model (Birnbaum, 1968), while the polytomous items were calibrated using generalized partial credit model (GPCM) (Muraki, 1992).

3.2 Item Response Theory Model: Generalized Partial Credit Model

Because the 2PL model is a special case of GPCM, this study treated the 2PL-calibrated parameters as a simplified instance of GPCM. GPCM describes the probability of an examinee with a latent trait level θ to obtain a score of $v \in \{0, 1, \ldots, m_i\}$ for item i as:

$$p_{iv} = \frac{\exp(\sum_{r=0}^{v} Da_i(\theta - b_i + d_{ir}))}{\sum_{c=0}^{m_i} \exp(\sum_{r=0}^{c} Da_i(\theta - b_i + d_{ir}))},$$

where a_i and b_i respectively denote the overall discrimination and difficulty parameters for item i, and d_{ir} represents the step parameter for the category r for the item. The GPCM parameters used in this study were estimated under the constraints $d_{i0} = 0$ and $\sum_r d_{ir} = 0$. For the difficulty modeling, the target variable was defined as the item category threshold $b_i - d_{ir}$ for polytomously scored items, where $d_{ir} = 0$ for the dichotomous cases.[2] The discrimination parameter a_i was used as the target variable for modeling item discrimination.

3.3 Sampling

This study adopted an 80%:10%:10% split approach to create training, development, and test sets, respectively. The training set was used to train the models, while the development set was used to guide hyperparameter tuning and modeling decisions. The test set was held out to ensure the generalized performance of the trained models. For the sampling, items were stratified by item types to equally distribute all item types across the sets. A detailed breakdown of item type counts across the three sets is provided in Appendix D.

3.4 Fine-tuning LLMs Using Texts as Primary Input

To evaluate how item texts can be fine-tuned for predicting IRT parameters, we implemented two distinct LLM architectures as shown in Appendix E: a baseline model that uses only item texts as input, and an experimental model that incorporates both item texts and attributes as input. In both architectures, the model started by encoding item text into static token-level embeddings of 768 dimensions using a pre-trained LLM. These token embeddings were then aggregated to a single 768-dimensional vector using mean pooling. Subsequently, the pooled vector was passed through three consecutive hidden layers, with each normalized by batch normalization, followed by Leaky ReLU activations (Maas et al., 2013). Finally, a regression head was attached to the last hidden layer to produce a continuous output for the target IRT parameters.

Model with Item Attributes. As shown on the right side of Figure 1, the experimental model with item attributes was implemented by concatenating item attribute variables with the pooled embeddings before passing it through the hidden layers. This design allowed the model to leverage both item text and attributes seamlessly within the fine-tuning process.

[1] Although the text input consists of rubric, item and stimulus texts, we refer to this combined input as "item text" for brevity throughout the remainder of this paper.

[2] These category threshold parameters are referred to as *difficulty* parameters throughout the remainder of this paper for simplicity.

Within this architecture, we explored two model variants: one that uses the raw item attribute variables directly for the concatenation and the other that concatenates predicted values from another predictive ML model based on item attributes. We refer to the first variant as the *feature augmented* model, where the raw variables are used to augment the LLM feature space. The other variant is referred to as the *transfer learning* model, as it transfers predictive outputs from a separate model into the LLM. Note that the inclusion of two additional model variants resulted in a total of three fine-tuned LLM approaches: (i) *baseline* models using item texts as the sole input, (ii) *feature augmented* models using raw item attributes alongside the texts, and (iii) *transfer learning* models that used LLM-predicted values as an additional input.

Selected Pre-trained LLMs. Four different LLMs were experimented in this study to evaluate differences in performance: RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), XLNet (Yang et al., 2019), and Longformer (Beltagy et al., 2020). Three models except for XLNet were encoder-based model. These encoder models were chosen as they are designed to transform texts into contextualized embeddings, which can be seamlessly adjusted for regression tasks. XLNet, while not an encoder-only model, was chosen as it often outperformed other transformer-based models due to its recurrent mechanism that can accommodate long-term dependencies (Ormerod et al., 2023).

Hyperparmeters. The following hyperparameter settings were used throughout this study:

- **Batch size:** 16 for most models; reduced to 8 for DeBERTa due to GPU memory limitations
- **Number of epochs:** 40 for most models; increased to 100 for DeBERTa to ensure sufficient updates to mitigate noisier gradients due to the smaller batch size [3]
- **Sequence length:** 512 [4]
- **Learning rate:** $1e^{-5}$ for pre-trained LLM parameters and $1e^{-2}$ for other model parameters

3.5 Traditional ML Approach Using Item Attributes as Primary Input: CatBoost

In addition to fine-tuned LLMs, this study implemented a traditional ML approach to predict IRT parameters using item attribute variables as primary input. Specifically, CatBoost (Dorogush et al., 2018)—a gradient boosting algorithm based on decision trees—was chosen due to its ability to natively handle categorical variables without requiring dummy coding [5]. Following the structure used in the LLM-based modeling, three variants of the CatBoost approach were developed: (i) *baseline* models, (ii) *feature augmented* models, and (iii) *transfer learning* models.

In the *baseline* model, only the item attribute variables were used as input features. For the *feature augmented* model, embeddings extracted from each of the fine-tuned baseline LLMs were appended to the item attribute feature set. In the *transfer learning* model, the predicted values generated by the fine-tuned baseline LLMs were appended to the item attribute feature set as an additional predictor.

4 Results

The predictive performance of the models was evaluated using two metrics: Pearson correlation (COR) and root mean squared error (RMSE). These metrics were calculated by treating the banked IRT parameters as the ground truth for the development and test sets. Table 1 displays descriptive statistics of the banked IRT parameters across the subsets.

Parameter	Statistic	Train	Dev	Test
Discrimination (a_i)	Min	0.110	0.166	0.203
	1st Qu.	0.448	0.437	0.466
	Median	0.579	0.548	0.633
	Mean	0.586	0.553	0.610
	3rd Qu.	0.718	0.681	0.743
	Max	1.354	1.043	1.075
	SD	0.199	0.179	0.209
Difficulty ($b_i - d_{ir}$)	Min	-2.719	-1.770	-1.631
	1st Qu.	-0.175	-0.024	-0.086
	Median	0.798	0.863	0.648
	Mean	0.891	0.978	0.805
	3rd Qu.	1.766	1.770	1.681
	Max	9.068	6.251	4.607
	SD	1.379	1.391	1.257

Table 1: Summary statistics for the banked IRT parameters across the sets.

The prediction results presented in the following are all based on the held-out test set.

[3]The study used the model state after completing all training epochs as the final model.

[4]While Longformer and XLNet can process inputs longer than 512 tokens, the sequence length was fixed at 512 based on preliminary analysis showing no performance advantage from longer inputs.

[5]The study used CatBoost v1.2.7 with default hyperparameter settings.

4.1 Item Difficulty Prediction Results

Table 2 presents the performance of fine-tuned LLM and CatBoost models on the item difficulty prediction task.

Positive Impact of Item Attribute Integration in LLM Fine-Tuning. The results demonstrate promising prediction accuracy for baseline fine-tuned LLMs, achieving correlations close to 0.7 with Longformer and DeBERTa. These findings suggest that item text alone can contribute substantial information relevant to predicting difficulty parameters when leveraged through LLM fine-tuning. Further improvements were observed with the *feature augmented* models, where raw item attribute variables were integrated into the LLM fine-tuning. This method consistently (albeit marginally) outperformed the baseline across all four LLMs, yielding the highest correlations and lowest RMSEs in most of LLMs.[6] These results indicate that item attributes can provide additional information in prediting difficulty parameters. In contrast, the *transfer Learning* LLM models—where predicted values from CatBoost model were appended to inputs—suffered reduced performance. This indicates that incorporating predicted values from an external model may have simply added additional noise rather than signal, particularly when the predictions themselves were only moderately accurate (e.g., a correlation of 0.492 in this case).

Improved CatBoost Performance via LLM-Based Augmentation. The CatBoost Baseline model, which used only item attribute variables for predicting item difficulty, showed limited predictive power, with a correlation less than 0.5. However, its performance improved substantially when augmented with embeddings from fine-tuned LLMs, regardless of the LLM type. For example, augmenting item attributes with fine-tuned embeddings increased the correlation from .492 to .706 in the case of Longformer. A similar positive effect was observed with *transfer learning* model; when predicted values from fine-tuned LLMs were added as additional inputs, performance markedly improved over baseline.

While both the *feature augmented* and *transfer learning* CatBoost models showed notable gains, their performance remained short of the best results achieved by the fine-tuned LLMs—those augmented with raw item attribute variables.

[6]The magnitude of improvement was more pronounced in the development set results; see Appendix F.

4.2 Item Discrimination Prediction Results

Table 3 presents the item discrimination prediction performance of fine-tuned LLM and CatBoost models. The discrimination prediction results showed distinctively different patterns from the difficulty prediction.

Strong Performance of CatBoost for Discrimination Prediction. In contrast to the pattern observed in difficulty prediction, the baseline CatBoost model yielded the strongest performance for discrimination prediction among all conditions. As shown in Table 3, this baseline model, which used only item attribute variables, achieved the highest correlation (0.537) and the lowest RMSE (0.174), consistently outperforming all other model variants.

In comparison, the baseline fine-tuned LLM models performed notably worse than they had in the difficulty prediction task. When fine-tuned solely on item text, the LLMs produced correlation values as low as 0.310. Likely due to this low baseline performance, augmenting CatBoost with fine-tuned embeddings or LLM-based predictions resulted in noticeable drops in prediction accuracy. For instance, with RoBERTa, the addition of fine-tuned embeddings reduced the correlation from 0.537 (CatBoost baseline) to 0.396. This degradation further highlights the limited capacity of fine-tuned LLMs for modeling item discrimination.

Conversely, the contribution of item attribute variables to the fine-tuned LLMs was notable, leading to consistent performance gains. For example, Longformer's correlation improved from 0.337 in the baseline LLM to 0.425 with the addition of raw item attributes, and further increased to 0.477 when predictions from the CatBoost model were appended. This trend was consistent across all LLMs: the *transfer learning* fine-tuned LLM models always outperformed the baseline LLMs, often by substantial margins.

5 Discussion

In this study, we presented novel approaches for fine-tuning LLMs using item text to predict IRT parameters. Beyond the baseline model that relied solely on item text, we introduced structured methods for incorporating item attribute variables into the fine-tuning process to further enhance predictive performance. We also examined the use

Approach	Variant	Correlation				RMSE			
		RoBERTa	DeBERTa	XLNet	Long former	RoBERTa	DeBERTa	XLNet	Long former
Fine-tuned LLM	Baseline	0.633	0.691	0.686	0.691	1.029	0.910	0.964	0.967
	F.A.	**0.707**	**0.712**	**0.699**	**0.717**	**0.890**	**0.889**	**0.908**	0.966
	T.L.	0.563	0.644	0.547	0.609	1.070	0.980	1.095	1.069
CatBoost	Baseline	0.492	0.492	0.492	0.492	1.133	1.133	1.133	1.133
	F.A.	0.669	0.686	0.559	0.706	0.965	0.936	1.054	**0.907**
	T.L.	0.646	0.678	0.688	0.696	0.997	0.936	0.919	0.913

Table 2: Item difficulty prediction results on the test set using fine-tuned LLM and CatBoost models across three model variants. Within each LLM, **bold** marks the best performance and underline marks the second best. F.A.=Feature Augmented; T.L.=Transfer Learning. Corresponding results to the development set can be found in Appendix F.

Approach	Variant	Correlation				RMSE			
		RoBERTa	DeBERTa	XLNet	Long former	RoBERTa	DeBERTa	XLNet	Long former
Fine-tuned LLM	Baseline	0.394	0.310	0.334	0.337	0.196	0.205	0.218	0.200
	F.A.	0.348	0.320	0.324	0.425	0.203	0.201	0.208	0.196
	T.L.	0.465	0.495	0.470	0.477	0.186	0.195	0.276	0.183
CatBoost	Baseline	**0.537**	**0.537**	**0.537**	**0.537**	**0.174**	**0.174**	**0.174**	**0.174**
	F.A.	0.396	0.428	0.426	0.409	0.196	0.192	0.190	0.197
	T.L.	0.414	0.352	0.376	0.335	0.194	0.200	0.200	0.202

Table 3: Item discrimination prediction results on the test set using fine-tuned LLM and CatBoost models across three modeling variants. Within each LLM, **bold** marks the best performance and underline marks the second best. F.A. = Feature Augmented; T.L. = Transfer Learning. Corresponding results to the development set can be found in Appendix G.

.

of a traditional ML algorithm—CatBoost—using item attribute variables as primary inputs, and further investigated whether combining CatBoost with information derived from fine-tuned LLMs could improve prediction accuracy.

Performance of the suggested methods was evaluated using a large dataset of Grade 6 ELA assessment items. The dataset included a mix of dichotomously and polytomously scored items, offering a valuable opportunity to assess model performance on predicting multiple difficulty parameters in polytomous items. In addition, we also fully investigated the prediction performance of the item discrimination parameters, which has received limited attention in prior IRT parameter modeling research.

Our results suggested that predicting item difficulty parameters was a relatively more amenable modeling task, with several models achieving moderately high correlations. In contrast, predicting item discrimination parameters were found to be more challenging, consistently yielding lower performance. In particular, we found that the fine-tuned LLMs performed well in the difficulty prediction, but were susbtantially less effective for discrimination. This disparity indicates that item texts contain meaningful signals for modeling difficulty, but offer limited information in capturing item discrimination.

Interestingly, the traditional CatBoost model using only item attribute variables showed relatively strong performance in predicting discrimination parameters, achieving highest correlations and lowest RMSEs. This finding highlights the potential value of using structured item attribute features in modeling discrimination parameters and may offer useful direction to researchers and practitioners.

The study also explored the integration of two information sources—item text and item attributes—as inputs into the prediction models. This strategy showed mixed results. When the added information came from a strong predictive source, such as fine-tuned LLM derived values in the difficulty modeling, it considerably enhanced model performance. However, when the appended information had limited predictive quality, it often introduced

noise and reduced accuracies. These findings highlight both the promise and the risks of multi-source modeling: while combining signals can enhance prediction, it is crucial to assess the individual contribution of each source before integration.

Limitations

This study is not without limitations. First, although we partitioned the dataset into training, development, and test sets, we did not employ full cross-validation during hyperparameter tuning. As a result, model performance may have been somewhat sensitive to the specific data split that we used. Second, all hyperparameter settings were optimized based on the development set performance for the difficulty prediction task. These settings were then applied to the discrimination prediction without futher tuning. Given the distinct nature of the prediction targets, task-specific hyperparameter optimization—particularly for discrimination modeling using fine-tuned LLMs—could have yielded improved performance. Third, while several models achieved strong correlations for difficulty prediction, the overall predictive accuracy indicates considerable potential for future improvement. This reflects the inherent complexity of this task and highlights the need for continued research.

Future Work

In an effort to improve the alignment of predicted values with the true parameters, the authors conducted preliminary investigation of a sequential approach that incorporates predicted values as informative priors within a Bayesian estimation framework with small samples, as explored in Ulitzsch et al. (2025). Initial results suggest that incorporating a small response sample—as small as 50 examinees—can significantly improve estimation accuracy. However, a detailed discussion of this extension lies beyond the scope of the current study and will be addressed in future work. In addition, future research will explore the differential performance of the predictions across item types as a larger and more diverse sample of items becomes available. Such analysis is expected to provide practical insights for practitioners by illuminating conditions where fine-tuned LLMs are most effective in predicting IRT parameters.

References

Samah AlKhuzaey, 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.

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2021. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, 58(1):4–30.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2021. On the application of transformers for estimating the difficulty of multiple-choice questions from text. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–157, Online. Association for Computational Linguistics.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

Allan Birnbaum. 1968. Some latent trait models. *Statistical theories of mental test scores*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Wangyong Feng, Peter Tran, Hunter McNichols, Steven Sireci, and Andrew Lan. 2024. Using artificial intelligence to scale multiple choice math items. Presentation delivered at the Annual Conference of the Northeastern Educational Research Association, Trumbull, CT.

Steve Ferrara, Jeffrey T Steedle, and Roger S Frantz. 2022. Response demands of reading comprehension test items: A review of item difficulty modeling studies. *Applied Measurement in Education*, 35(3):237–253.

Gerhard H Fischer. 1995. The linear logistic test model. In *Rasch models: Foundations, recent developments, and applications*, pages 131–155. Springer.

Sebastian Gombert, Lukas Menzel, Daniele Di Mitri, and Hendrik Drachsler. 2024. Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 483–492.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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*.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, and 1 others. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA.

Eiji Muraki. 1992. A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, 16(2):159–176.

Christopher Ormerod, Amy Burkhardt, Mackenzie Young, and Sue Lottridge. 2023. Argumentation element annotation modeling using xlnet. *arXiv preprint arXiv:2311.06239*.

Georg Rasch. 1993. *Probabilistic models for some intelligence and attainment tests.* ERIC.

Esther Ulitzsch, Dmitry Belov, Oliver Lüdtke, and Alexander Robitzsch. 2025. Using item parameter predictions for reducing calibration sample requirements—a case study based on a high-stakes admission test. *Journal of Educational Measurement*.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pages 193–197.

Victoria Yaneva, Peter Baldwin, Janet Mee, and 1 others. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 11–20.

Victoria Yaneva, Peter Baldwin, Christopher Runyon, and 1 others. 2023. Extracting linguistic signal from item text and its application to modeling item characteristics. In *Advancing natural language processing in educational assessment*, pages 167–182. Routledge.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jiyun Zu and Ikkyu Choi. 2023. Predicting the psychometric properties of automatically generated items. Presentation delivered at the 88th Annual Meeting of the Psychometric Society, College Park, MD.

A Descriptions of Item Types

Abbreviation	Item Type	Description
EBSR	Evidence-Based Selected Response	This item type has two parts: Part A asks examinees to select a correct response from four options, and Part B asks them to identify textual support for their answer
HT	Hot Text	This item type asks examinees to either select a correct word or rearrange words/phrases by clicking and dragging
MC	Multiple Choice	This item type asks examinees to choose one answer from multiple options
MI	Match Interaction	This item type requires examinees to match text or images in rows to values in columns by clicking cells
MS	Multi Select	This item type asks examinees to select one or more options
SA	Short Answer	This item type asks examinees to enter a response using alphanumeric characters via a keyboard
WER	Writing Extended Response	This item type asks examinees to provide a longer written response using keyboard entry of alphanumeric characters

Table 4: Descriptions of the item types used in this study

B Item Counts by Year and Type

Year	EBSR	HT	MC	MI	MS	SA	WER	Total
2014	23	75	174	1	119	42	6	440
2015	20	79	127	1	85	31		343
2016		1	7		4			12
2017	1	2	8		3	2	3	19
2018	2	6	9	5	3	19	4	48
2019	11	24	84	4	14	1	2	140
2020						5		5
2022	3	16	55	4	11	21	2	112
Total	60	203	464	15	239	116	22	1119

Table 5: Item counts by field test year and item type

C Examples of Item Attribute Variables

Attribute Type	Variable Type	Label	Description
Content Spec	Categorical	itemType	Item types
Content Spec	Categorical	WERdimension	Dimension of WER items
Content Spec	Categorical	claim	Four main claims in Smarter Balanced ELA
Content Spec	Categorical	lowestLevel	Content standards for ELA
Content Spec	Categorical	stimGenre	Genre of stimulus
Content Spec	Numeric	IAT.Tables	Number of tables embedded in the item
Content Spec	Numeric	IAT.Images	Number of images embedded in the item
Content Spec	Numeric	choiceInt	Number of choice-type interactions in the item
Content Spec	Numeric	hotTextInt	Number of hot-text-type interactions in the item
Content Spec	Numeric	FleschEase	The Flesch Reading readability level measuring easiness of text
Content Spec	Numeric	FleschKinc	The Flesch Kincaid Readability level measuring US grade level required to understand text
Linguistic	Numeric	numWords	Number of words in the text
Linguistic	Numeric	numContWords	Number of content words in the text
Linguistic	Numeric	numPolySem	Number of words that have multiple meanings
Linguistic	Numeric	numWSenseNoun	Number of word senses for nouns
Linguistic	Numeric	avgSynTreeDep	Average depth of syntax trees in sentences
Linguistic	Numeric	notCommon2000	Number of words that are not in common 2000 words in Reuter corpus
Linguistic	Numeric	avgImage	Average rating of words based on how easily and quickly a mental image can be evoked, according to the MRC Psycholinguistic Database

Table 6: Example of item attribute features. Content Spec=Content-based specification features; Linguistic=Handcrafted linguistic features

D Distribution of Item Types Across Subsets

Item Type	Training	Development	Test	Total
EBSR	48	6	6	60
HT	162	20	21	203
MC	371	46	47	464
MI	12	1	2	15
MS	191	24	24	239
SA	92	12	12	116
WER	17	2	3	22
Total	893	111	115	1119

Table 7: Counts of Item Types by Subset.

E Fine-tuned LLM Model Architecture

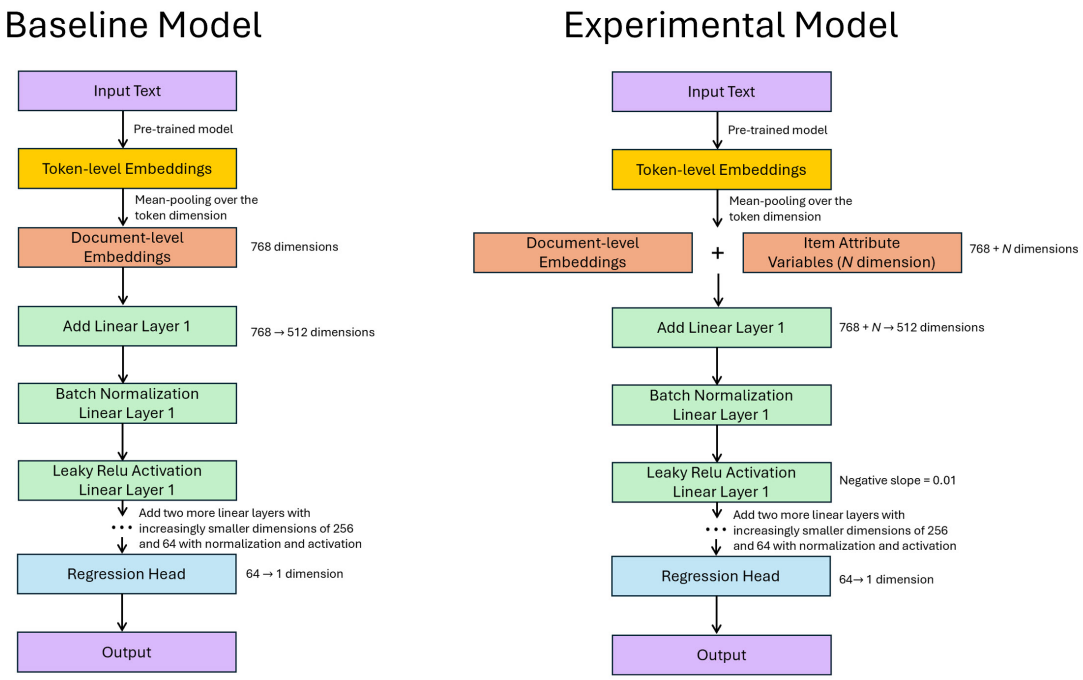

Figure 1: Fine-tuned LLM Model Architecture

F Item Difficulty Prediction Results on the Development Set

Approach	Variant	Correlation				RMSE			
		RoBERTa	DeBERTa	XLNet	Long former	RoBERTa	DeBERTa	XLNet	Long former
Fine-tuned LLM	Baseline	0.742	0.732	0.713	0.744	0.930	0.961	1.065	0.931
	F.A.	**0.762**	0.777	**0.777**	**0.785**	0.932	0.886	**0.884**	**0.882**
	T.L.	0.759	**0.792**	0.752	0.767	0.924	**0.867**	0.938	0.908
CatBoost	Baseline	0.692	0.692	0.692	0.692	1.008	1.008	1.008	1.008
	F.A.	0.729	0.728	0.749	0.772	0.953	0.970	0.932	0.896
	T.L.	0.754	0.742	0.741	0.751	**0.915**	0.944	0.952	0.927

Table 8: Item difficulty prediction results on the development set using fine-tuned LLM and CatBoost approaches across three model variants. Within each LLM, **bold** marks the best and underline marks the second best performance. F.A.=Feature Augmented; T.L.=Transfer Learning.

G Item Discrimination Prediction Results on the Development Set

Approach	Variant	Correlation				RMSE			
		RoBERTa	DeBERTa	XLNet	Long former	RoBERTa	DeBERTa	XLNet	Long former
Fine-tuned LLM	Baseline	0.258	0.216	0.190	0.188	0.187	0.192	0.204	0.250
	F.A.	0.374	0.241	0.292	0.371	<u>0.180</u>	0.186	0.191	<u>0.174</u>
	T.L.	<u>0.406</u>	<u>0.339</u>	<u>0.368</u>	<u>0.386</u>	0.183	<u>0.175</u>	0.316	0.178
CatBoost	Baseline	**0.447**	**0.447**	**0.447**	**0.447**	**0.167**	**0.167**	**0.167**	**0.167**
	F.A.	0.272	0.235	0.296	0.294	0.186	0.194	<u>0.190</u>	0.185
	T.L.	0.304	0.221	0.221	0.266	0.184	0.192	0.196	0.186

Table 9: Item discrimination prediction results on the development set using fine-tuned LLM and CatBoost approaches across three model variants. Within each LLM, **bold** marks the best and <u>underline</u> marks the second best performance. F.A.=Feature Augmented; T.L.=Transfer Learning.