

Multimodal Item Parameter Estimation using Simulated Response Probabilities

Christopher Ormerod
College Board
cormerod@collegeboard.org

YoungKoung Kim
College Board
ykim@collegeboard.org

Abstract

We present results from reconstructing multiple-choice model (MCM) and three-parameter logistic (3PL) model curves using a fine-tuned multimodal large language model (LLM) based on Qwen3.5. The model is prompted and fine-tuned to replicate choice probabilities across a large training corpus of multiple-choice items containing both image and text stimuli, conditioned on a labeled set of student ability levels. By learning to reproduce the systematic error patterns of students across a discrete range of abilities, the LLM implicitly captures the underlying response probabilities encoded in the 3PL and MCM curves. This allows us to accurately approximate item difficulty on a held-out test set directly from the model's predicted option probabilities.

1 Introduction

Item difficulty parameters are essential for evaluating whether an assessment accurately measures student knowledge, differentiating between students, and identifying specific learning gaps. The field-testing required to obtain these parameters is not only expensive and labor-intensive, but also poses an inherent security risk. For these reasons, researchers have explored machine learning techniques to approximate these parameters (AlKhuyaey et al., 2024). Given the capabilities of Large Language Models (LLMs), a growing body of literature has emerged on their use in predicting item difficulty parameters. Two characteristically distinct approaches have been proposed: treating the question stimulus and its features as input to a fine-tuned LLM in a regression framework (Li et al., 2025), and using LLMs as simulated respondents to test stimuli (Maeda, 2025).

Generative LLMs have become increasingly powerful due to advances in training regimes (Wang et al., 2024) and the ability to scale the transformer architecture (Vaswani et al., 2017) to

the point at which emergent capabilities arise (Wei et al., 2022). The dominant paradigm for generative models has shifted from open-ended text generation to instruction-following via prompting (Chung et al., 2022). The flexibility of prompting allows researchers to instruct LLMs to answer questions in targeted ways, enabling the simulation of students with specified combinations of skills (Lu and Wang, 2024), knowledge (He-Yueya et al., 2024), and ability levels (Liu et al., 2025). The present study extends work in which models are fine-tuned in a parameter-efficient manner to simulate students of varying ability (Scarlatos et al., 2025), subsequently applied to multiple-choice questions (Ormerod, 2026).

Prior work has examined the reconstruction of item characteristic curves (Ormerod, 2026) using the Nominal Response Model (NRM) and the Two-Parameter Logistic (2PL) model to estimate item parameters (Thissen and Steinberg, 1986). Both models share the limiting assumption that the probability of a correct response approaches zero as ability decreases. The Three-Parameter Logistic (3PL) model (Thissen and Steinberg, 1986) and the Multiple-Choice Model (MCM) (Thissen and Steinberg, 1984; Samejima, 1969) address this by incorporating non-zero lower asymptotes, thereby accounting for the possibility of guessing. First, the present work introduces several improvements to the framework of (Ormerod, 2026) and applies it to a large corpus of mathematics items. Second, we demonstrate this framework for items with both text and images by fine-tuning multimodal models with hybrid transformer-based architectures. While this may appear rather cumbersome, we are able to reproduce the item difficulty parameter with remarkably high accuracy compared to baseline regression-based methods.

We organize this paper as follows: Our method section, §2, covers the nature of the response and item data used in this project, the response model-

ing applied to determine item parameters, the way in which we discretize ability levels, and the way we use fine-tuned generative LLMs, the baseline approaches, and the metrics used for evaluation. In §3, we present our results, before concluding with a discussion in §4.

2 Method

To explain our method, we have raw data in the form of responses, and various models based on that data: 3PL models (Thissen and Steinberg, 1986), MCMs (Thissen and Steinberg, 1984), the discrete MCMs (Ormerod, 2026), and LLMs (Touvron et al., 2023). These models have been represented in Figure 1, while the arrows, labeled i to vii , denote the various ways to define parameters of each model either in terms of the raw responses or other models.

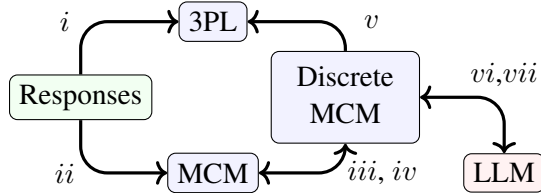


Figure 1: Diagram of the relationships between the data and models used in this study.

The following subsections are devoted to defining the data used in this study, the models, and the method of deriving the parameters of each model. With all these defined, our goal is to reconstruct the parameters of the 3PL models and MCMs defined by i and ii , respectively.

2.1 Data

The raw data in this study consists of two distinct objects: items and responses to those items.

2.1.1 Items

Each item consists of a stimulus and a set of four distinct options. Since our understanding of Mathematics can be inherently visual, we allow both the stimulus and the options to have an associated image component. In order to facilitate this combination of images as input, we transform the images into one image in the manner presented in Figure 2.

This dataset is produced using item models, which specify the structure, content, and variable features needed to produce new, functional test items (Gierl and Haladyna, 2013). There were a

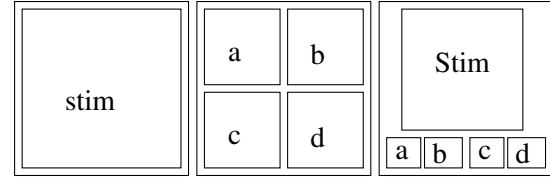


Figure 2: The way in which stimulus and option images were combined into a single image. The left image depicts stimulus-only images, the middle depicts option-only images, and the right depicts when both stimulus and options are present.

	Item Model	Items	Responses
Train	727	3633	10.2×10^6
Development	97	485	1.42×10^6
Test	146	730	2.24×10^6
Total	970	4848	13.88×10^6

Table 1: A representation of how the items were split into test, train, and development.

total of 970 item models and a total of 4,848 items. In order to avoid conflation, we split the items into training, development, and test sets by item model. This gave a natural split at the item level represented in Table 1.

2.2 Response Modeling

There are 13.88 million responses. The ability level of each respondent was independently derived from a larger calibrated dataset of items. These responses provide us with an average of approximately 2860 responses per item with a minimum of 860 responses. The average response ability level was 0.31 and the standard deviation was approximately 1.

Given that each of these responses can be associated with a particular value of student ability, we can model the probability that a student j , with ability level θ_j , provides a correct answer to item i by

$$P_{ij}(X = 1) = c_i + (1 - c_i)\sigma(a_i(\theta_j - b_i)) \quad (1)$$

where σ is the usual sigmoid function. There are at least 10 correct or incorrect responses to any particular item, facilitating stable estimates of the item parameters, $\{a_i, b_i, c_i\}$ by minimizing the negative log likelihood function using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno with Bound constraints (L-BFGS-B).

To satisfy the conditions that $a_i > 0$ and $0 < c_i < 1$, we let $a_i = e^{\alpha_i}$ and $c_i = \sigma(\gamma_i)$. Then our first arrow, i , is provided by the optimization

$$a_i, b_i, c_i = \operatorname{argmax} \left\{ \sum_j [y_j \log P_{ij} + (1 - y_j) \log(1 - P_{ij})] - \lambda(\alpha_i^2 + b_i^2) \right\} \quad (i)$$

where the first two terms are the usual negative log likelihood function and the third term is a small $L2$ regularization term with $\lambda = 10^{-4}$. We used the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) method.

The model described by (1) is known as the three-parameter logistic model (3PL) (Thissen and Steinberg, 1986). We call a_i the discrimination parameter, b_i the difficulty parameter, and c_i is called the guessing parameter. It has the asymptotic property that as θ_j is large and negative, the probability of obtaining the correct response is given by random chance; hence, it is typically assumed to be close to the reciprocal of the number of options. As $c_i \rightarrow 0$, we obtain the two-parameter logistic model (Thissen and Steinberg, 1986).

The 3PL model is useful in defining a difficulty parameter; however, it does not allow us to provide any details regarding the strength of distractor items. Given options $k = 1, \dots, K$, the probability of a person of ability level θ_j providing answer k to item i is modeled by

$$P_{ijk} = P_{ij}(X = k) = \frac{\exp(\eta_{ijk}) + d_{ik}}{1 + \sum_k \exp(\eta_{ijk})}, \quad (2)$$

where $\eta_{ijk} = a_{ik}\theta_j + b_{ik}$. Given that $\sum_k P_{ij}(X = k) = 1$ and $d_{ik} > 0$, this provides us with $\sum_k d_{ik} = 1$ for any item. This model is a reformulation of the multiple-choice model from Thissen and Steinberg (1984). Alternatively, this generalizes the version of Samejima in which $d_{ik} = K^{-1}$ (1969). In a similar manner to the 3PL model, we may derive the values of $\{a_{ik}, b_{ik}, d_{ik}\}$ using the optimization

$$\{a_{ik}, b_{ik}, d_{ik}\} = \operatorname{argmax} \left\{ \sum_{j=1} \log P_{ijk} - \lambda(a_{ik}^2 + b_{ik}^2) \right\} \quad (ii)$$

with $\lambda = 10^{-4}$ using LBFGS. This provides us with the arrow ii . In combination with i , this provides the target values for our modeling.

2.3 Discrete Ability Modeling

The main idea from (Ormerod, 2026) is that we are binning the values of θ_j into L categories. The cutoff points for each category are denoted γ_i so that $I_l = (\gamma_{l-1}, \gamma_l]$ with

$$-\infty = \gamma_0 < \gamma_1 < \dots < \gamma_L = \infty.$$

This provides us with intervals $I_1, \dots, I_L$ such that each $\theta_j \in I_{l_j}$. The greater the L , the more fine-grained the division of abilities. Each interval is associated with a loosely descriptive label. The labels and intervals are presented in Table 2.

Label	γ_{i-1}	γ_i
Foundational	$-\infty$	-2.7
Beginning	-2.7	-2.4
Emerging foundations	-2.4	-2.1
Early developing	-2.1	-1.8
Developing	-1.8	-1.5
Developing proficiency	-1.5	-1.2
Approaching basic	-1.2	-0.9
Basic	-0.9	-0.6
Basic plus	-0.6	-0.3
Approaching average	-0.3	0
Average	0	0.3
Average plus	0.3	0.6
Above average	0.6	0.9
Strong	0.9	1.2
Very strong	1.2	1.5
Advanced	1.5	1.8
Highly advanced	1.8	2.1
Exceptional	2.1	2.4
Outstanding	2.4	2.7
Elite mastery	2.7	∞

Table 2: A list of the descriptive labels used to calibrate the language model.

Following the work of (Ormerod, 2026), for item i , we associate each interval with a probability value for each option, which gives us a function

$$f_i(I_j) = (\rho_{ij1}, \dots, \rho_{ijK}). \quad (3)$$

This means we obtain, as data, an $L \times K$ matrix of probability values.

Our underlying prior is that the population is normally distributed such that $\{\theta_j\} \sim \mathcal{N}(\mu, \varsigma^2)$. We find it convenient to state functions associated with this distribution explicitly. The probability and

cumulative distribution functions (PDF & CDF), $f(\theta)$ and $F(\theta)$ are given by

$$f(\theta) = \frac{1}{\varsigma} \phi\left(\frac{\theta - \mu}{\varsigma}\right), \quad F(\theta) = \Phi\left(\frac{\theta - \mu}{\varsigma}\right),$$

where

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad \Phi(z) = \int_{-\infty}^z \phi(t) dt,$$

It makes sense to define the conditional probability that $\theta \in I_j$ as

$$w_j = F(\gamma_j) - F(\gamma_{j-1})$$

In (Ormerod, 2026), each interval was associated with the expected value under this prior, with ρ_{ijk} defined as the evaluation of (2) at this value. It actually makes more sense to associate the interval with the expected value of the function itself. Provided that P_{ijk} is defined by (2), then the definition of ρ_{ijk} may be determined by

$$\rho_{ijk} = \frac{1}{w_j} \int_{\gamma_{j-1}}^{\gamma_j} f(\theta) P_{i\theta k} d\theta, \quad (iii)$$

where $P_{i\theta k}$ is the function of θ rather than the particular value at θ_j . This provides us with *iii*.

Furthermore, given that these are continuous values rather than discrete values, we can approximate the reconstruction of the MCM by optimizing

$$(a_{ik}, b_{ik}, d_{ik}) = \operatorname{argmin} \left(\sum_j w_j \left(\rho_{ijk} - \frac{1}{w_j} \int_{\gamma_{j-1}}^{\gamma_j} f(\theta) P_{i\theta k} d\theta \right)^2 \right). \quad (iv)$$

Alternatively, we can specialize the value of k to be the correct option, $\tilde{k}$, in which case we obtain the probability of being correct, which provides a viable path to reconstructing the parameters of (1). For completeness, we write this as

$$(a_i, b_i, c_i) = \operatorname{argmin} \left(\sum_j w_j \left(\rho_{ij\tilde{k}} - \frac{1}{w_j} \int_{\gamma_{j-1}}^{\gamma_j} f(\theta) P_{i\theta} d\theta \right)^2 \right) \quad (v)$$

where $P_{i\theta}$ is the model (1) where the dependency on θ_j is replaced by the general function of θ . This provides us with all the arrows that do not depend on the language model.

It is worth noting that these approximations, at each arrow, are not exact. For example, the arrow i provides a mapping from the set of empirically observed responses to the IRT parameters; however, there is an underlying assumption that (1) is a good fit for the empirical data. We assume that there are some regressive tendencies in which the parameters are linearly related to the final parameters. We use the development set to appropriately and linearly model this relationship, which does not change the overall Pearson correlations, but does affect the mean squared error.

2.4 Language Modeling

The Qwen model series continues to provide the research community with a suite of excellent fine-tunable models for research purposes. The latest series of models, Qwen3.5, provides us with the opportunity to demonstrate the abilities of multimodal fine-tuning. The models we conservatively chose are the 4-billion- and 9-billion-parameter variants of Qwen3.5.

2.4.1 Parameter-efficient fine-tuning

These models are small enough to fit on a local machine with sufficient resources. In order to train these models, we employ parameter-efficient methods (Xu et al., 2023) such as Low-Rank Adaptation (LoRA) (Hu et al., 2021) or LoRA with Quantization (Dettmers et al., 2023) to effectively tune the model, requiring some knowledge of the underlying model structure.

The Qwen3.5 series not only integrates multimodal inputs but is also a hybrid structure similar to the Jamba series (Lieber et al., 2024). Each layer replaces multiheaded attention in the transformer architecture (Vaswani et al., 2017) with two distinct types of layers: Gated DeltaNet layers (Yang et al., 2025) and Gated Attention layers (Qiu et al., 2025). The Gated DeltaNet layers have a structure similar to that of the Mamba2 layers in that the information is stored in hidden states and preserved through a selective gating mechanism. This means that the Gated DeltaNet layers have linear complexity while the Gated Attention layers have quadratic complexity, making this closer to a Jamba model than previous Qwen models.

Figure 4 presents the internal structure of the Gated DeltaNet and Attention layers.

The application of LoRA takes a selection of linear layers, of the form

$$L(x) = Mx + b$$

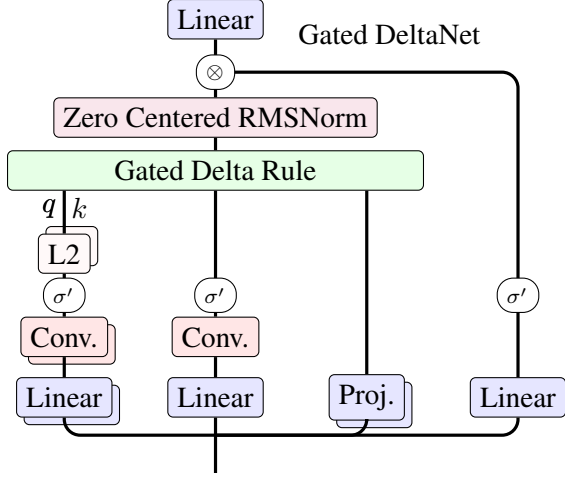


Figure 3: The structure of the Gated DeltaNet layers. Here, σ' is the SiLU function.

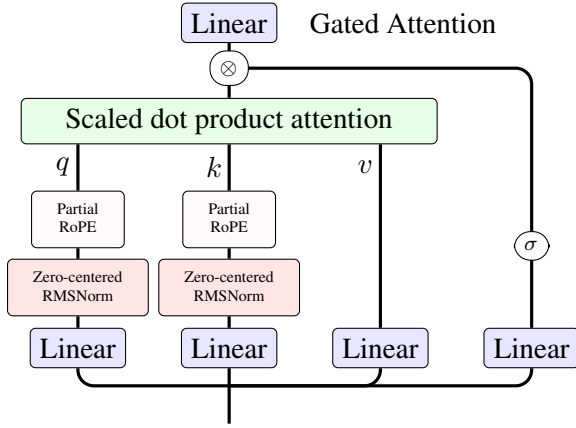


Figure 4: The structure of the Gated Attention layers. Here, σ is the sigmoid function.

where M is an $m \times n$ matrix, and replaces those linear layers with

$$\tilde{L}(x) = (M + BA)x + b$$

where A is an $r \times n$ matrix and B is an $m \times r$ matrix with $r \ll \min(m, n)$. The main idea is that we freeze all other layers and keep A and B trainable (Hu et al., 2021). This process has been very successful in automated scoring (Ormerod and Kwako, 2024), standards alignment (Han et al., 2026), and item parameter estimation (Ormerod, 2026).

Given instabilities in the training process encountered in the literature, we conservatively applied LoRA to the attention components, knowing this would essentially affect only a quarter of the hidden layers that would be typically used to fine-tune models. We would need a better understanding of how to fine-tune the Gated DeltaNet layers before considering the other layers.

2.4.2 Prompting

The creation of a dataset that encodes information regarding student options requires formatting the data as a sequence of chats. Each chat consists of three components: a system text indicating the behavior of the model, a user text indicating a task to be completed, and an assistant text indicating how the task should be completed. These texts are functions of the ability level (`ability_level`), the correct answer (`correct`), the item text (`stim_text`), and the option text (`option_text`), which contains an enumeration of the options in text form. These are presented below:

System: You are a student who must answer either A, B, C, D in a multiple choice question.

Ability Level: {ability_level}.

Correct Answer: {correct}

User: Answer the following **Question** as if you were a {ability_level} student.

Question
{stim_text}

Options
{option_text}

Assistant: **Answer**:

We leave the final text of the student option blank, truncating the text so that it does not satisfy the traditional requirements of a full chat. The idea is that we are obtaining a vector of next-token probabilities, which we will denote v where the dimension of v is the number of tokens in the model vocabulary. For convenience, we will assume the softmax has been applied so that each component of v , denoted v_i , is interpreted as a token probability.

There are indices, $\iota_1, \dots, \iota_K$, associated with each option. This allows us to construct the Mean

Squared Error (MSE) loss function, defined by

$$\mathcal{L} = \sum_k \frac{1}{K} (v_{\iota_k} - \rho_{ijk})^2, \quad (4)$$

where i denotes the item index and j denotes the ability level index. Using the standard Adam optimizer with a weight decay mechanism and a learning rate of 10^{-5} over 10 epochs provides us with the training regime used to define the fine-tuned language model, essentially providing the arrow vi . Finally, inference using the prompt above provides the arrow vii .

2.4.3 Baseline models

Our baseline approach is to use the question stimulus and options as input and then treat the IRT parameters as targets for a regression problem. This means that we take an encoder-only transformer-based language model, such as MetaMath (Yu et al., 2024) and MathBERT (Peng et al., 2021), replace the head with a linear layer with one target, and use the MSE loss function, similar to above, with the outputs of the language model and the target IRT parameters as the arguments. This approach is the one that is most common in the literature (Bulut et al., 2024; Peters et al., 2025).

Table 3 summarizes the models used in this study, together with their approximate parameter counts and references.

2.5 Metrics

In reference to the literature on item parameter prediction, there are two dominant metrics used to compare the performance of difficulty prediction models: Pearson correlation and root mean squared error (AlKhuyaey et al., 2024).

We would like to highlight another approach that fits nicely in the context of our study: a measure of the agreement under a discretization of difficulty. In communicating difficulty to educators, item difficulty parameters are often difficult to interpret. For this reason, what is often conveyed is a descriptive label that is similar to the descriptive labels we used to describe ability levels. The labels, as functions of the b parameter, are presented in Table 4. Under these descriptive labels, it is useful to obtain a measure of the agreement between the predicted labels and the calibrated labels. For this, we use the quadratic weighted kappa (QWK).

3 Results

3.1 Difficulty Prediction

Difficulty is the parameter of greatest practical interest, since it is the quantity most often reported to educators and used when assembling test forms. Table 5 reports the Pearson correlation and root mean squared error (RMSE) between the predicted and calibrated parameters.

The fine-tuned Qwen3.5-9B model attains a Pearson correlation of 0.85 on the difficulty parameter b , exceeding both regression baselines—MathBERT (0.68) and MetaMath (0.75)—by a clear margin. This corresponds to a relative improvement of roughly 15% over the stronger baseline, obtained with an RMSE of 0.55 after the linear development-set correction described in Section 2. The result indicates that the simulated-respondent framing recovers difficulty-relevant signal that the direct stimulus-to-parameter regression of the baselines does not capture.

When difficulty is discretized into the five-band scheme of Table 4, the model reaches a QWK of 0.835 against the calibrated labels, whereas MetaMath has a QWK of 0.692 and MathBERT has a QWK of 0.625.

4 Discussion

Item difficulty modeling remains one of the most difficult tasks in computational psychometrics, and our results reinforce both the promise and the limits of the simulated-respondent approach. With a modest amount of parameter-efficient fine-tuning, a 9-billion-parameter multimodal model reconstructs the difficulty parameter at a fidelity (Pearson 0.85, QWK 0.835) that comfortably exceeds strong text-based regression baselines, and it does so for items that mix textual and visual stimuli within a single model and a single input representation (Figure 2). This suggests that the central idea—training the model to replicate the systematic pattern of option choices across a discretized range of abilities and then reading item parameters off the resulting curves—transfers cleanly to the multimodal setting without changes to the underlying architecture.

The most informative aspect of our results is the uneven recovery across parameters. Recovering the guessing parameter c at a correlation of 0.48, where the regression baselines are very low by comparison, is arguably the clearest evidence that the approach captures genuine response behavior rather than surface features of the stimulus: a

Model	Role	Params	Reference
MathBERT	Baseline	110M	(Peng et al., 2021)
MetaMath	Baseline	70B	(Yu et al., 2024)
Qwen3.5-4B	Proposed	4B	(Qwen Team, 2026)
Qwen3.5-9B	Proposed	9B	(Qwen Team, 2026)

Table 3: Models used in this study, with approximate parameter counts and references. MathBERT (Peng et al., 2021) is built on BERT-base; MetaMath is a LLaMA-2 fine-tune available at 7B, 13B, and 70B.

Label	Full Name	Condition
VH	Very Hard	$b \geq 0.842$
H	Hard	$0.253 \leq b < 0.842$
M	Medium	$-0.253 \leq b < 0.253$
E	Easy	$-0.842 \leq b < -0.253$
VE	Very Easy	$b < -0.842$

Table 4: A list of the descriptive labels used to communicate difficulty.

	Pearson		
	a	b	c
MathBERT	0.35	0.68	0.25
MetaMath	0.34	0.75	0.28
Qwen3.5-4B	0.26	0.80	0.45
Qwen3.5-9B	0.31	0.85	0.48
	RMSE		
	a	b	c
MathBERT	1.03	0.78	0.12
MetaMath	1.04	0.68	0.12
Qwen3.5-4B	1.01	0.63	0.10
Qwen3.5-9B	0.85	0.55	0.09

Table 5: The associated correlation values and RMSE values for each approach.

model that merely reads the question cannot easily infer how often low-ability students will be drawn to a particular distractor, whereas a model trained to imitate those students can. Conversely, the weak recovery of the discrimination parameter a indicates that slope information is partially lost, most plausibly through the discretization of ability into intervals and through the regression-based correction applied to the raw estimates. Finer-grained ability binning, at additional computational cost, is a natural avenue for improving slope recovery.

Several design choices bound these results and point to concrete extensions. First, to avoid the training instabilities reported in the literature, we applied LoRA only to the Gated Attention compo-

nents of the hybrid Qwen3.5 architecture, leaving the Gated DeltaNet layers untouched; this adapts roughly a quarter of the layers that a conventional transformer would expose to fine-tuning. A better understanding of how to adapt the DeltaNet layers stably would likely improve recovery, particularly of the parameters that are weakest under the present regime. Second, every arrow in our pipeline introduces approximation error—the mapping from empirical responses to IRT parameters assumes that the 3PL and MCM are good fits, and the discrete reconstruction replaces continuous curves with binned expectations—so the reported correlations are best read as lower bounds on what the framing can achieve under tighter modeling assumptions.

Acknowledgments

References

- Samah AlKhuyaey, 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.
- Okan Bulut, Guher Gorgun, and Bin Tan. 2024. [Item Difficulty and Response Time Prediction with Large Language Models: An Empirical Analysis of USMLE Items](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 522–527, Mexico City, Mexico. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. [Scaling Instruction-Finetuned Language Models](#). *arXiv preprint*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient Fine-tuning of Quantized LLMs](#). *Advances in Neural Information Processing Systems*, 36:10088–10115.

- Mark J. Gierl and Thomas M. Haladyna, editors. 2013. *Automatic Item Generation: Theory and Practice*. Routledge.
- Suhwa Han, Christopher Ormerod, and Frank Rijmen. 2026. *Fine-tuning of Open-source Large Language Models to Align Items to Standards*. Preprint, OSF/EdArXiv.
- Joy He-Yueya, Wanjing Anya Ma, Kanishk Gandhi, Benjamin W. Domingue, Emma Brunskill, and Noah D. Goodman. 2024. *Psychometric Alignment: Capturing Human Knowledge Distributions via Language Models*. *arXiv preprint*. ArXiv:2407.15645 [cs.CL].
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *LoRA: Low-Rank Adaptation of Large Language Models*. *arXiv preprint*.
- 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*. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 48–55, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirum, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, and 3 others. 2024. *Jamba: A Hybrid Transformer-Mamba Language Model*. *arXiv preprint*. ArXiv:2403.19887 [cs.CL].
- Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2025. *Leveraging LLM respondents for item evaluation: A psychometric analysis*. *British Journal of Educational Technology*, 56(3):1028–1052. *_eprint*: <https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13570>
- Xinyi Lu and Xu Wang. 2024. *Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation*. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pages 16–27. ArXiv:2405.11591 [cs.HC].
- Hotaka Maeda. 2025. *Field-Testing Multiple-Choice Questions With AI Examinees: English Grammar Items*. *Educational and Psychological Measurement*, 85(2):221–244.
- Chris Ormerod and Alexander Kwako. 2024. *Automated Text Scoring in the Age of Generative AI for the GPU-poor*. *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3).
- Christopher Ormerod. 2026. *Reconstructing Item Characteristic Curves using Fine-Tuned Large Language Models*.
- Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. *MathBERT: A Pre-Trained Model for Mathematical Formula Understanding*. *arXiv preprint*.
- Sydney Peters, Nan Zhang, Hong Jiao, Ming Li, Tianyi Zhou, and Robert Lissitz. 2025. *Text-Based Approaches to Item Difficulty Modeling in Large-Scale Assessments: A Systematic Review*.
- Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. *Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free*. *arXiv preprint*. ArXiv:2505.06708 [cs.CL].
- Qwen Team. 2026. *Qwen3.5: Towards Native Multi-modal Agents*.
- Fumiko Samejima. 1969. *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. *Psychometrika*, 34(S1):1–97.
- 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 25071–25094, Suzhou, China. Association for Computational Linguistics.
- David Thissen and Lynne Steinberg. 1984. *A response model for multiple choice items*. *Psychometrika*, 49(4):501–519.
- David Thissen and Lynne Steinberg. 1986. *A taxonomy of item response models*. *Psychometrika*, 51(4):567–577.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. *arXiv preprint*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is All you Need*. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. 2024. *Reinforcement Learning Enhanced LLMs: A Survey*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. *Emergent Abilities of Large Language Models*.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. [Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment](#). *arXiv preprint*.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. 2025. [Gated Delta Networks: Improving Mamba2 with Delta Rule](#). *arXiv preprint*. ArXiv:2412.06464 [cs.CL].

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [Meta-Math: Bootstrap Your Own Mathematical Questions for Large Language Models](#). *arXiv preprint*. ArXiv:2309.12284 [cs.CL].