DEEP-MIRT: Dynamic Psychometric Knowledge Tracing for Partial-Credit and Rubric-Scored Assessment

Wenrui Yuan, Stéphanie van den Berg, Johannes Steinrücke

February 20, 2026

Abstract

Knowledge tracing is a core component of adaptive learning systems, yet much of the literature assumes binary correctness as the observation signal. In many educational settings, however, learner evidence is polytomous, including ordered rubric levels and partial credit, where both ordinal structure and interpretability matter for measurement and intervention. We present DEEP-MIRT, a modular framework that extends knowledge tracing to polytomous responses by bridging sequential neural modeling with psychometric Item Response Theory. The framework couples a swappable sequential backbone with an interchangeable IRT-grounded response head, separating the question of how learner ability evolves over time from the question of how that ability maps to ordered response categories. Instantiated with a Generalized Partial Credit Model (GPCM) head, DEEP-MIRT produces interpretable outputs at every interaction: an evolving student ability estimate, per-item discrimination values, and ordered difficulty thresholds corresponding to rubric level transitions. On synthetic benchmarks with known ground truth, DEEP-GPCM achieves strong rank-order agreement across category resolutions $K \in \{2, 3, 4, 5\}$ (QWK up to 0.778) and recovers item difficulty thresholds and discrimination with high to moderate fidelity (mean $r_\beta = 0.90$; $r_\alpha \approx 0.65$). The model remains backward-compatible with standard binary knowledge tracing benchmarks, matching established baselines on ASSIST2015 and Synthetic-5 with $K = 2$. Together, these properties advance knowledge tracing from a binary predictive instrument into a measurement-aware framework that supports both instructional decision-making and principled psychometric interpretation in partial-credit and rubric-scored assessment contexts.

Introduction

Educational assessments often target graded performance rather than pass/fail outcomes. Rubric-based scoring, partial-credit items, and constructed-response tasks encode degrees of correctness that binary labels collapse into a single signal (Black and Wiliam, 1998; Heritage, 2007). While automated scoring systems can deliver this ordinal structure at scale in intelligent tutoring contexts (Shute, 2011; Baral et al., 2021), most KT benchmarks, evaluation protocols, and dominant modeling paradigms still assume binary correctness. This leaves ordinal outcomes comparatively underexplored in KT, even though they are central to authentic assessment practice.

Beyond the outcome type, interpretability remains a key concern. Most KT models learn high-dimensional states optimized for prediction, but these states are not, by construction, tied to interpretable quantities. Therefore, constructs that educators and assessment designers routinely work with, such as student ability level, item discriminatory power, or the difficulty steps between adjacent rubric categories, are not directly identifiable. Item Response Theory (IRT) provides exactly this structure through identifiable parameters including student ability, item discrimination, and ordered step difficulties, grounding score interpretation in established psychometric theory (Muraki, 1997; Embretson and Reise, 2000; Messick, 1995). Standard IRT models, however, treat ability as fixed and are not designed for sequential interaction data. This work therefore extends KT to polytomous responses by incorporating the interpretable parameter structure of IRT within a dynamic sequential framework.

Three lines of existing work address these issues in isolation, leaving a gap at their intersection. First, deep KT models, from Deep Knowledge Tracing (DKT; Piech et al., 2015) to memory-augmented and attention-based variants (Zhang et al., 2017; Pandey and Karypis, 2019), capture temporal learning dynamics but typically assume dichotomous correctness and learn latent states with limited psychometric interpretability. Second, polytomous Item Response Theory (IRT) models such as the Generalized Partial Credit Model (GPCM; Muraki, 1997) provide principled, interpretable measurement of ordered responses with identifiable parameters (student ability, item discrimination, ordered step thresholds), but they generally treat ability as time-invariant and are not designed for sequential interaction logs. Third, recent efforts that inject IRT structure into deep KT (Yeung, 2019; Converse et al., 2021) improve interpretability but largely remain restricted to dichotomous outcomes. An early partial exception is Wang and Heffernan (2013), who extended Bayesian Knowledge Tracing to allow partial credit via continuous Bayesian nodes, but without deep sequential modeling or recovering standard polytomous IRT step structure. This leaves open the question of how to unify dynamic sequential modeling with ordinal response modeling in a framework that exposes psychometrically identifiable parameters.

We propose *DEEP-MIRT*, a modular framework that extends KT to Polytomous Knowledge Tracing (PKT) by coupling a swappable sequential backbone with an IRT-grounded response head. In this setting, PKT is the task of predicting the full distribution over ordered response categories from a learner's interaction history. We instantiate the framework with a GPCM response head, yielding DEEP-GPCM, which combines a Dynamic Key-Value Memory Network (DKVMN) backbone with a GPCM observation model. At each timestep, the GPCM head maps the learner state to a distribution over ordinal categories and exposes psychometrically interpretable components, including an evolving student ability estimate, as well as learned item parameters such as per-item discrimination values and ordered step difficulties corresponding to rubric level transitions. While the experiments focus on ordinal PKT, the framework accommodates other polytomous formats through alternative response heads.

Our main contributions are:

1. **A modular framework for polytomous knowledge tracing.** We formulate PKT as predicting the full ordinal response distribution from interaction history and propose DEEP-MIRT, which decouples a sequential backbone (encoder, memory, summary) from an interchangeable psychometric response head. The head interface accepts student ability θ_t and item identity $\mathbf{k}_t$ and returns a category probability vector; any differentiable observation model satisfying this contract is a valid instantiation. This design separates the educational measurement question (which model governs how ability maps to rubric scores?) from the knowledge tracing question (how does ability evolve over time?).

2. **GPCM instantiation with interpretable psychometric parameters.** We instantiate DEEP-MIRT with a GPCM-based response head as DEEP-GPCM, incorporating a monotonic threshold parameterization that enforces rubric ordering by construction. The resulting model produces evolving ability trajectories, item discrimination estimates, and ordered step difficulty thresholds, all of which correspond to constructs that educators and assessment designers already use in rubric design and standard-setting practice.

3. **Scalable item encoding and empirical evaluation.** We propose Static Item Embedding (SIE) for large item banks and evaluate ordinal prediction quality, psychometric parameter recovery, ablation of design components, and backward compatibility with binary KT benchmarks.

We investigate the following research questions:

- **RQ1 (Prediction).** Does coupling a DKVMN backbone with an ordinal, psychometrically-constrained GPCM response head improve polytomous next-response prediction, and does it better preserve ordinal structure than format-unaware alternatives?

- **RQ2 (Parameter Recovery).** Do the IRT parameters inferred by DEEP-GPCM faithfully recover known ground-truth values from synthetic data, and are item parameter estimates stable across student subpopulations as assessed by split-half reliability?

- **RQ3 (Learner State).** Do the inferred student ability trajectories θ_t exhibit psychologically plausible learning dynamics, and can DEEP-GPCM reveal meaningful learner patterns (mastery, growth, persistent difficulty) that are not visible from format-unaware latent state representations?

- **RQ4 (Scalability).** Does Static Item Embedding maintain ordinal prediction quality and parameter recovery at large item-bank scales ($Q \sim 5{,}000$) while substantially reducing parameter count compared to dense encoding strategies?

- **RQ5 (Ecological Validity).** When ordinal response categories are constructed from proxy signals in existing binary KT benchmarks (e.g., attempt counts), does the ordinal GPCM head yield measurable gains over format-unaware alternatives, and how sensitive are these gains to the quality of the proxy ordinal construction?

Related Work

Knowledge Tracing: From Binary to Polytomous

Knowledge tracing has evolved through several modeling paradigms, but all share a common assumption: learner responses are binary. Bayesian Knowledge Tracing (BKT) models mastery of a Knowledge Component (KC) as a two-state Hidden Markov Model (Corbett and Anderson, 1994). Deep Knowledge Tracing (DKT) reframed the task as sequence modeling with recurrent neural networks, learning high-dimensional latent states from interaction sequences (Piech et al., 2015). Dynamic Key-Value Memory Networks (DKVMN) introduced a structured memory that separates static concept representations (key memory) from evolving mastery states (value memory), offering a more transparent mapping between knowledge components and latent state than generic RNNs (Graves et al., 2016; Zhang et al., 2017). Attention-based models such as SAKT and AKT apply self-attention to weigh the relevance of prior interactions for next-step prediction (Pandey and Karypis, 2019; Ghosh et al., 2020), and recent work has further explored Transformer-based architectures (Choi et al., 2020; Liu et al., 2023). Throughout this progression, the response format has remained dichotomous: every model cited above predicts a binary correct/incorrect outcome.

Recent KT research has begun to acknowledge that correctness alone is insufficient. Lu et al. (2024) fuse process features (attempts, response time, historical accuracy) to enrich difficulty representations, and Li et al. (2025) integrate problem complexity and state-stability mechanisms to reduce noisy updates. These efforts treat richer signals as auxiliary features within a binary prediction framework. They do not change the response format itself, but they highlight that educational interactions carry information beyond pass/fail: precisely the information that ordinal and polytomous responses encode directly. The only prior work to move beyond binary responses in a KT context is Wang and Heffernan (2013), who extended BKT to partial credit using continuous Bayesian nodes, but without deep sequential modeling or IRT-based parameterization.

Interpretability in Deep Knowledge Tracing

In most DKT models, learned states are opaque distributed representations rather than psychometrically meaningful constructs (National Research Council, 2001). Several efforts address this gap by coupling neural architectures with Item Response Theory (IRT) structure. Deep-IRT attaches an IRT-style readout to DKVMN, exposing student ability and item difficulty parameters for each prediction (Yeung, 2019). DKVMN-MRI similarly integrates IRT within an LSTM-based memory framework (Xu et al., 2024). Converse et al. (2021) propose a general approach for incorporating IRT parameters into KT training. These IRT-augmented models improve interpretability: a teacher or assessment specialist can inspect the estimated ability and difficulty values, rather than relying on opaque hidden states. In educational terms, "interpretability" here means that model outputs correspond to constructs that practitioners already use (student proficiency levels, item difficulty, and discrimination quality), supporting the kinds of validity arguments that responsible assessment requires (Messick, 1995; Kane, 2013).

A critical limitation, however, is that all existing IRT-KT integrations operate on dichotomous responses. They estimate a single ability and difficulty for binary outcomes and do not model the ordered step structure of rubric scores or partial-credit items. Attention-based models provide a complementary form of interpretability through attention weights (which interactions were most relevant to a prediction), but they too lack psychometric parameterization and ordinal response awareness.

Polytomous Item-Response Theory

Psychometrics provides a family of models for ordered responses grounded in a latent ability parameter θ and item-specific properties (Hambleton and Swaminathan, 1985; Embretson and Reise, 2000). Two traditions dominate for ordinal outcomes. Cumulative models such as the Graded Response Model (GRM; Samejima, 1969) define boundary parameters between adjacent categories, modelling the probability of scoring at or above category k as a logistic function of θ. Adjacent-category models, including the Partial Credit Model (PCM; Masters and Wright, 1997) and the Generalized Partial Credit Model (GPCM; Muraki, 1997), parameterize each step transition directly: the GPCM characterizes each item by a discrimination parameter α_j controlling how sharply the item differentiates ability levels, and ordered step difficulty thresholds $\boldsymbol{\beta}_j = (\beta_{j,0}, \dots, \beta_{j,K-2})$ representing the ability level at which each consecutive rubric transition becomes equally likely. Adjacent-category models are the natural fit for rubric scoring, where performance progresses sequentially through ordered levels (Embretson and Reise, 2000; Wilson, 2005). For unordered categories, the Nominal Response Model (NRM; Bock, 1972) provides an analogous framework.

All classical polytomous IRT models treat ability as fixed across items, which is appropriate for standardized testing but inadequate for adaptive tutoring contexts where ability evolves through practice. Extending the interpretable parameter structure of IRT to a dynamic sequential setting is the central goal of this work.

Methodology

Framework Overview

DEEP-MIRT is a modular architecture with two separable concerns. The *backbone* handles the knowledge tracing problem: it encodes item–response interactions, maintains a dynamic memory of the student's evolving knowledge state, and produces a summary vector $\mathbf{f}_t$ representing current student state. The *response head* handles the measurement problem: it maps the backbone's student state $\mathbf{f}_t$ and item identity $\mathbf{k}_t$ to a probability distribution over response categories through a psychometric observation model. Any differentiable function $h(\mathbf{f}_t, \mathbf{k}_t) \to \Delta^{K-1}$ (a distribution over K categories) satisfying this interface is a valid DEEP-MIRT instantiation. The backbone is agnostic to which measurement model governs the head.

This separation matters educationally. The backbone addresses the learning dynamics question of how student ability evolves across a sequence of interactions, best handled by sequential neural architectures. The head addresses the measurement question

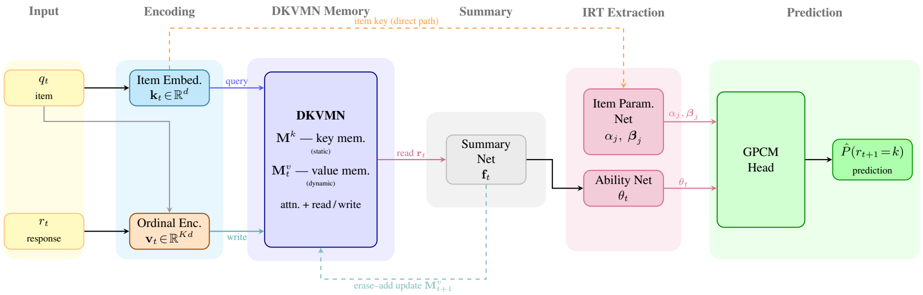

Figure 1: DEEP-MIRT framework at timestep t. Item parameters $(\alpha_j, \boldsymbol{\beta}_j)$ bypass the student state via the item embedding (dashed, top), consistent with the IRT assumption that discrimination and difficulty are student-independent. The GPCM head maps $(\theta_t, \alpha_j, \boldsymbol{\beta}_j)$ to the predicted ordinal distribution $\hat{P}(r_{t+1} = k)$.

of how current ability maps to the probability of each ordered rubric score, best handled by psychometric theory. Coupling them through a clean interface means the two concerns can be improved independently: a better backbone (e.g., Transformer instead of DKVMN) improves temporal modeling without changing the interpretable parameterization; a different head (e.g., GRM instead of GPCM) changes the measurement assumptions without retraining the backbone.

In this paper we instantiate the head as the **Generalized Partial Credit Model** (GPCM, Section). We chose GPCM over the other major polytomous families for three reasons: (1) it is an adjacent-category model, which is the natural match for rubric scoring where a student progresses sequentially through performance levels; (2) it produces a discrimination parameter α_j and ordered step difficulty thresholds $\boldsymbol{\beta}_j$ that correspond directly to rubric design decisions and can be inspected by assessment practitioners; (3) its monotone threshold constraint admits a clean differentiable parameterization (Section). A cumulative model such as the Graded Response Model (Samejima, 1969) would be a natural alternative for tasks where each scoring threshold represents an independent decision (e.g., analytic rubrics with independent sub-scores); a Nominal Response Model (Bock, 1972) would apply when response categories are unordered (e.g., multiple-choice distractors encoding different misconceptions). We treat these as directions for future instantiation.

The backbone processes inputs through four sequential stages visible in Figure 1: ordinal interaction encoding, DKVMN memory update, summary compression, and IRT-based parameter extraction. An ordinal-aware training objective supervises all stages jointly. The following subsections formalize each component.

Problem Formulation

Let $r_t \in \{0, 1, \ldots, K-1\}$ denote an observed response with K *ordered* categories (e.g., rubric levels / partial credit), and let $q_t \in \{0,1\}^Q$ be a one-hot indicator of the exercised question/item, where Q is the number of unique items. An interaction is $x_t = (q_t, r_t)$ and the observed history up to time t is $\mathcal{X}_t = \{(q_1, r_1), \ldots, (q_t, r_t)\}$. We study *ordinal* (polytomous) knowledge tracing as predicting the full distribution

$$P(r_{t+1} = k \mid q_{t+1}, \mathcal{X}_t), \quad k \in \{0, \ldots, K-1\}. \tag{1}$$

Ordinal-Aware Interaction Representation

Triangular One-Hot Encoding We encode an ordinal response r_t with a "triangular" kernel over ordered categories. Define a length-K ordinal code $\mathbf{s}_t \in \mathbb{R}^K$ by

$$\begin{aligned} s_t^{(k)} &= \max\left(0, 1 - \frac{|k - r_t|}{K-1}\right), \quad 0 \le k \le K-1, \\ \mathbf{s}_t &= [s_t^{(0)}; \ldots; s_t^{(K-1)}]. \end{aligned} \tag{2}$$

For example, with a $K = 4$ rubric and a response of $r_t = 2$ ("correct approach with errors"), the ordinal code is $\mathbf{s}_t = [0.33,\ 0.67,\ 1.0,\ 0.67]$, which spreads activation to neighboring categories while peaking at the observed level. This encoding preserves the ordinal structure: a response of 2 is represented as more similar to responses 1 and 3 than to response 0, reflecting the educational intuition that partial credit is closer to full credit than to no credit.

We then bind the ordinal code to the item indicator via a Kronecker product (equivalently, a block one-hot expansion):

$$\mathbf{x}_t = q_t \otimes \mathbf{s}_t \in \mathbb{R}^{KQ}. \tag{3}$$

This reduces to the standard binary interaction encoding when $K = 2$ (consistent with Zhang et al. (2017)). While this encoding is effective for moderately sized item banks, the $\mathcal{O}(KQ)$ dimensionality becomes a practical constraint for large Q; we address this in the following subsection.

Scalable Value Encoding for Large Item Banks Large-scale educational platforms routinely maintain item banks with thousands of items across courses and grade levels. The triangular Kronecker encoding in (2) yields an interaction vector of dimension KQ, which grows linearly with item bank size. The subsequent value projection $\mathbf{v}_t = B^\top \mathbf{x}_t \in \mathbb{R}^{d_v}$ therefore requires a $KQ \times d_v$ weight matrix; at $Q = 5{,}000$ and $K = 5$, this amounts to 1.6M parameters and a per-batch intermediate tensor of size $\text{batch} \times S \times 25{,}000$, placing substantial memory pressure on GPU-constrained deployment environments.

We propose **Static Item Embedding** (SIE), an alternative value encoding that decouples item identity from response pattern via two independent additive branches:

$$\mathbf{v}_t = W_{\text{item}}\, \mathbf{u}_{q_t} + W_{\text{resp}}\, \mathbf{s}_t + \mathbf{b}_{\text{resp}} \in \mathbb{R}^{d_v}, \tag{4}$$

where $\mathbf{u}_{q_t} \in \mathbb{R}^H$ is a *frozen* (non-trainable) unit-norm random vector assigned to item q_t at initialization, $\mathbf{s}_t \in \mathbb{R}^K$ is the triangular ordinal code from (2), and $W_{\text{item}} \in \mathbb{R}^{d_v \times H}$ and $W_{\text{resp}} \in \mathbb{R}^{d_v \times K}$ are learned projections shared across all items.

The factored structure in (4) separates two gradient channels: W_{item} is updated by item-identity signals, while W_{resp} is updated by response-pattern signals. Because item vectors are frozen, items cannot collapse to similar representations during training, a failure mode of learned separable embeddings at large Q (cf. Section). For two independent uniformly random unit vectors in $\mathbb{R}^H$, the expected squared inner product is $1/H$ (Dasgupta and Gupta, 2003), so pairwise interference concentrates around zero as H grows; cross-item gradient crosstalk in W_{item} therefore decays with H independently of Q, making SIE well-suited for large item banks.

The frozen dimension H is set as the smallest power of two no less than $\max(128, \lfloor Q/2 \rfloor)$, clamped to $[128, 1024]$.[1] This schedule ensures near-orthogonality scales with item bank size while keeping the projection dimension bounded, in contrast to the $\mathcal{O}(KQ)$ complexity of the full Kronecker encoding.

Dynamic Key-Value Memory Network

We adopt the Dynamic Key-Value Memory Network (DKVMN) as the sequential backbone. Let $\mathbf{k}_t = A^\top q_t \in \mathbb{R}^{d_k}$ be a key embedding for the current item and $\mathbf{v}_t \in \mathbb{R}^{d_v}$ be a value embedding for the interaction, obtained via the Kronecker projection $\mathbf{v}_t = B^\top \mathbf{x}_t$ (ordinal-aware encoding) or the Static Item Embedding in (4) (scalable encoding). The model maintains a static key memory $\mathbf{M}^k \in \mathbb{R}^{N \times d_k}$ and a dynamic value memory $\mathbf{M}^v_t \in \mathbb{R}^{N \times d_v}$. Educationally, the key memory represents latent skill areas or knowledge components, the value memory tracks the student's evolving mastery of each area, and the attention weights $\mathbf{w}_t$ (below) indicate which skill areas a given item exercises. The summary vector $\mathbf{f}_t$ aggregates the student's current mastery state relevant to the item at hand.

Specifically, the key query is first projected via a learned transform, and attention weights over memory slots are computed as

$$\tilde{\mathbf{k}}_t = \tanh(W_{\text{attn}} \mathbf{k}_t), \qquad \mathbf{w}_t = \operatorname{softmax}(\mathbf{M}^k \tilde{\mathbf{k}}_t) \in \mathbb{R}^N, \qquad w_t(i) = [\mathbf{w}_t]_i, \tag{5}$$

and the read vector is

$$\mathbf{r}_t = \sum_{i=1}^{N} w_t(i)\, \mathbf{M}^v_t(i) \in \mathbb{R}^{d_v}. \tag{6}$$

We fuse item context and retrieved state into a summary vector

$$\mathbf{f}_t = \tanh(W_{\mathbf{f}}[\mathbf{r}_t; \mathbf{k}_t] + b_{\mathbf{f}}), \tag{7}$$

which is then used by the response head to predict r_{t+1}.

Following memory-augmented neural networks, the value memory is updated via erase-add operations:

$$\mathbf{e}_t = \sigma(W_{\mathbf{e}} \mathbf{v}_t + b_{\mathbf{e}}), \tag{8}$$

$$\mathbf{a}_t = \tanh(W_{\mathbf{a}} \mathbf{v}_t + b_{\mathbf{a}}), \tag{9}$$

$$\mathbf{M}^v_{t+1}(i) = \mathbf{M}^v_t(i)\big[1 - w_t(i)\mathbf{e}_t\big] + w_t(i)\mathbf{a}_t. \tag{10}$$

[1] Concretely, $H = 128$ for $Q \leq 256$, $H = 256$ for $Q \approx 500$, $H = 512$ for $Q \approx 1{,}000$, and $H = 1{,}024$ for $Q \geq 2{,}000$.

DEEP-GPCM: an Ordinal IRT Head

Instead of the binary readout (e.g., $\sigma(W_p^\top \mathbf{f}_t + b_p)$), we use a Generalized Partial Credit Model (GPCM) head to obtain category probabilities for ordered responses. Let j denote the item exercised at time t (i.e., $[q_t]_j = 1$). We compute a time-varying ability proxy θ_t and item-specific parameters (thresholds and discrimination) as

$$\theta_t = W_\theta \mathbf{f}_t + b_\theta, \tag{11}$$

$$\beta_{j,0} = W_\beta \mathbf{k}_t + b_\beta, \tag{12}$$

$$\beta_{j,k} = \beta_{j,k-1} + \mathrm{softplus}(W_{g[k]}\mathbf{k}_t + b_{g[k]}), \quad 1 \le k \le K-2, \tag{13}$$

$$\alpha_j = \exp(W_\alpha \mathbf{k}_t + b_\alpha)\,. \tag{14}$$

The construction in (13) enforces $\beta_{j,0} < \beta_{j,1} < \cdots < \beta_{j,K-2}$ by adding strictly positive increments (via softplus) to each preceding threshold. This monotonicity guarantee has direct educational meaning: achieving rubric level $k{+}1$ always requires higher ability than achieving level k, respecting the ordering that a rubric designer intended. Without this constraint, the model could learn disordered thresholds where some rubric levels are never the most probable response at any ability level, a well-documented problem in educational measurement that signals a poorly functioning item (Wilson, 2005).

Because discrimination is an item-intrinsic property in the GPCM, α_j is derived solely from the item key $\mathbf{k}_t$, deliberately excluding the student-state summary $\mathbf{f}_t$. This separation ensures that the estimated discrimination reflects the item's inherent capacity to differentiate ability levels, independently of which student is currently being modelled — consistent with the IRT assumption that item parameters are population-invariant (Hambleton and Swaminathan, 1985). Item-level discrimination is recovered by averaging per-timestep estimates across all encounters with a given item (Meulders and Xie, 2004). The exponential activation ensures $\alpha_j > 0$.

Given $(\theta_t, \alpha_j, \{\beta_{j,k}\}_{k=0}^{K-2})$, the GPCM category probability is

$$p_{t,k} = \frac{\exp\Big(\sum_{i=0}^{k-1} \alpha_j(\theta_t - \beta_{j,i})\Big)}{\sum_{c=0}^{K-1} \exp\Big(\sum_{i=0}^{c-1} \alpha_j(\theta_t - \beta_{j,i})\Big)}. \tag{15}$$

At each timestep, the model evaluates this standard GPCM formulation (cf. Section) using the *dynamically estimated* student ability θ_t and item parameters $\alpha_j, \boldsymbol{\beta}_j$. Unlike classical GPCM estimation where ability is static, here θ_t evolves as new interactions are observed, enabling the model to track learning over time while retaining the interpretable measurement structure of polytomous IRT.

We predict $\hat{r}_{t+1} = \arg\max_k p_{t,k}$, and optimize the combined ordinal loss in (16) over the observed responses.

Training Objective for Ordinal Data

Because the GPCM defines category probabilities via (15) rather than cumulative links, we optimize a weighted categorical objective with an ordinal distance penalty to mitigate class imbalance (Gutiérrez et al., 2016). The loss combines two complementary components:

$$\mathcal{L} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{WOL}}\mathcal{L}_{\text{WOL}}, \tag{16}$$

where $\mathcal{L}_{\text{focal}}$ is the focal loss (Lin et al., 2020), which down-weights well-classified examples and focuses training on ambiguous responses near category boundaries. The weighted ordinal loss $\mathcal{L}_{\text{WOL}}$ is

$$\mathcal{L}_{\text{WOL}} = \frac{1}{n}\sum_{i=1}^{n}\Big[(1+\gamma|y_i - \hat{y}_i|)\cdot w_{y_i}\cdot(-\log \hat{p}_{i,y_i})\Big], \tag{17}$$

where y_i is the true category, $\hat{p}_{i,y_i}$ is the predicted probability under (15), $\hat{y}_i = \arg\max_k \hat{p}_{i,k}$ is the predicted label, and γ controls the ordinal distance penalty. To balance classes, with c_k the sample count of class k, we set $w_k = \sqrt{\frac{n}{K\cdot c_k}}$, where n is the number of valid tokens in the batch.

Experiments

Experimental Settings

Unless otherwise stated, all results below use **single-trait** instantiations of DEEP-MIRT; multidimensional extensions are part of future work.

Datasets

Synthetic-Ordinal. To the best of our knowledge, there are no standardized public benchmarks for ordinal knowledge tracing. We therefore generate a synthetic dataset using an IRT-GPCM data generating process following established psychometric principles. This design provides controlled polytomous supervision $r_t \in \{0,\ldots,K-1\}$ and enables parameter recovery analysis with known ground truth.

We simulate $5{,}000$ students interacting with questions drawn from a pool of $Q = 200$ questions in random order. Each interaction sequence has variable length in $[50, 200]$. Students and questions are configured with *static* GPCM parameters following assumed priors:

- *Student ability:* $\theta \sim \mathcal{N}(0,1)$,
- *Item discrimination:* $\alpha \sim \mathrm{LogNormal}(0, 0.3)$,
- *Item thresholds:* $K-1$ values drawn independently from $\mathcal{N}(0,1)$ and sorted to obtain $\beta_0 < \cdots < \beta_{K-2}$.

We generate four datasets with $K \in \{2,3,4,5\}$ to test extensibility across category resolutions. Each dataset is split 80/20 for training/testing; a validation split is held out from training for early stopping and hyperparameter tuning.

Proxy-Ordinality To test our model on real educational logs, we derive ordinal labels from attempt counts in ASSISTments-style data (e.g., 1 attempt = mastery, 2–3 = partial, 4+ = struggle). Results and evaluation details are described in Section .

Binary Compatibility After validating the PKT instantiations on Synthetic-Ordinal, we confirm compatibility with binary KT. Due to implementation differences in batch handling and sequence-to-question lookup, we select two datasets:

- **ASSIST2015**[2]: 19,840 students and 100 KCs/questions.
- **Synthetic-5**[3]: (Piech et al., 2015) simulated 4,000 students' responses with probability $p(r) = c + \frac{1}{1+\exp(-(\theta-\beta))}$ where $c = 0.25$ models question guessing. Each student is administered 50 questions in sequential order.

Models and Baselines

We compare DEEP-GPCM against three baselines that progressively relax psychometric and ordinal structure, each isolating a different design question:

- **DKVMN + Softmax**: DKVMN backbone with a K-way softmax head trained with categorical cross-entropy. No ordinal structure, no IRT constraints. Tests whether psychometric parameterization provides gains over a purely neural polytomous baseline.
- **DKVMN + Ordinal**: DKVMN backbone with a linear output head trained using the weighted ordinal loss (17). Introduces ordinal awareness without IRT parameterization, isolating the contribution of the GPCM structure over simpler ordinal objectives.
- **Static GPCM**: Standard GPCM fit from interaction data with student ability held fixed across items. No sequential dynamics. Represents the psychometric status quo and quantifies what sequential modeling adds beyond classical IRT estimation.
- **DEEP-GPCM (ours)**: DKVMN backbone with GPCM head and combined ordinal loss (16). Uses SIE encoding by default; the choice of encoding strategy is ablated in Section .

For dichotomous KT ($K = 2$), we additionally compare against **DKT** (Piech et al., 2015) and **DKVMN** (Zhang et al., 2017) as established binary baselines.

Prediction Mapping

For $K = 2$, predicted probabilities are converted to labels with a threshold of 0.5 following Zhang et al. (2017). For $K > 2$, we use $\hat{y}_t = \arg\max_k \hat{p}_{t,k}$ to map categorical probabilities to prediction labels.

[2] https://sites.google.com/site/assistmentsdata/datasets/2015-assistments-skill-builder-data

[3] https://github.com/chrispiech/DeepKnowledgeTracing/tree/master/data/synthetic

Training Setup

We use the combined loss in (16) to handle class imbalance and ordinal agreement, with tunable weights λ_*.

We set $d_k = 64$ and $d_v = 128$ for key/value embedding dimensions and configure $N = 50$ memory slots for both memory matrices.

All models are trained with batch size 64 using Adam with learning rate 10^{-3}, reduced on performance plateau as per Mukherjee et al. (2019) for late-epoch stability. We apply dropout 0.05 and early stopping if validation loss does not improve for 10 epochs. Models are implemented in PyTorch and trained on an RTX 4060 Laptop GPU.

Evaluation Metrics

We use metrics chosen to reflect ordinal prediction quality rather than mere categorical accuracy:

- **Categorical Accuracy (ACC)**: percentage of exact matches between predicted category and ground truth. While straightforward, ACC does not distinguish near-misses from far-misses and is therefore insufficient as a sole ordinal metric.

- **Quadratic Weighted Kappa (QWK)** (Cohen, 1968): our primary metric for ordinal tasks. QWK applies a quadratic penalty proportional to the squared distance between predicted and true categories, making it the standard agreement measure in automated essay scoring and rubric evaluation. A QWK of 0.0 indicates chance-level agreement; 1.0 indicates perfect agreement.

- **Kendall's Tau ($\boldsymbol{\tau}$)**: measures rank-order agreement between predicted and true response sequences, penalizing all inversions equally regardless of distance.

- **Area Under the ROC Curve (AUC)**: used for the binary case ($K = 2$) to maintain comparability with prior KT literature.

IRT Analysis Setup

For parameter recovery analysis on Synthetic-Ordinal, we use:

- **$\boldsymbol{\theta}$ Aggregation**: we take the mean ability estimate $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ across all timesteps for each student, averaging out sequence-position noise while exploiting the full interaction history.

- **$\boldsymbol{\alpha}, \boldsymbol{\beta}$ Aggregation**: because questions appear with variable frequency due to random sequencing, we aggregate item parameters based on occurrences and use the sample mean for recovery analysis.

We normalize learned parameters and true parameters using the priors used to generate each dataset. Specifically, both θ and β use z-score normalization. α is log-transformed and standardized with log-mean and log-standard deviation; we then rescale by 0.3 before the final exponential projection to match the generating scale.

Results

Ordinal Prediction on Synthetic-Ordinal (RQ1)

Table 1 reports prediction performance on Synthetic-Ordinal for $K \in \{2, 3, 4, 5\}$, addressing **RQ1**. DEEP-GPCM results are complete; baseline results are pending implementation (see note below). For DEEP-GPCM, ordinal agreement improves with category resolution: QWK rises from 0.618 at $K = 3$ to 0.758 at $K = 5$, and Kendall's τ from 0.567 to 0.655, indicating that the GPCM head effectively exploits the finer ordinal structure available in higher-K rubrics.

Table 1: Synthetic-Ordinal results across $K \in \{2, 3, 4, 5\}$ on Synthetic-Ordinal ($Q = 200$). Dashes indicate pending experiments. Best completed results in bold.

(a) $K = 2$ (Binary)

Model	Acc(%)	AUC	τ
DKT	—	—	—
DKVMN	—	—	—
DEEP-GPCM	**71.4**	**0.795**	**0.434**

(b) $K = 3$

Model	Acc(%)	QWK	τ
DKVMN + Softmax	—	—	—
DKVMN + Ordinal	—	—	—
Static GPCM	—	—	—
DEEP-GPCM	**60.7**	**0.618**	**0.567**

(c) $K = 4$

Model	Acc(%)	QWK	τ
DKVMN + Softmax	—	—	—
DKVMN + Ordinal	—	—	—
Static GPCM	—	—	—
DEEP-GPCM	**53.7**	**0.710**	**0.627**

(d) $K = 5$

Model	Acc(%)	QWK	τ
DKVMN + Softmax	—	—	—
DKVMN + Ordinal	—	—	—
Static GPCM	—	—	—
DEEP-GPCM	**48.7**	**0.758**	**0.655**

Pending: baseline experiments for RQ1

Three baselines remain to be implemented and evaluated on Synthetic-Ordinal.

- ***DKVMN + Softmax***: *Expected to achieve competitive accuracy but lower QWK and τ than DEEP-GPCM, since treating rubric levels as unordered categories ignores ordinal distance in prediction.*
- ***DKVMN + Ordinal***: *Expected to close part of the QWK gap over Softmax by penalising distant misclassifications. Whether it matches DEEP-GPCM will test whether the GPCM parameterization adds value beyond the ordinal loss alone.*
- ***Static GPCM***: *Expected to lag dynamic models substantially, particularly on longer sequences, as it cannot adapt ability estimates across interactions. Recovery of α and β may be comparatively clean, providing a useful psychometric reference point.*

Implementation note: DKVMN + Softmax and DKVMN + Ordinal share the backbone and differ only in the output head and loss; Static GPCM requires a standard marginal maximum likelihood fitting routine (e.g., EM with normal-ogive approximation).

Binary Compatibility on Standard KT Benchmarks

An important practical requirement is that a polytomous KT model should not sacrifice performance when deployed on standard binary data, since many existing item banks and datasets use dichotomous scoring, and a useful framework must handle both formats without architectural modification. We evaluate backward compatibility by instantiating DEEP-MIRT with $K = 2$ and testing on ASSIST2015 and Synthetic-5. Results are reported in Table 2. DEEP-GPCM achieves competitive performance compared to DKT and DKVMN despite substituting the binary readout with a GPCM head, confirming that the GPCM head gracefully degenerates to a standard logistic model when $K = 2$.

Table 2: Test results on selected metrics for ASSIST2015 and Synthetic-5. The best results for each metric are highlighted in bold.

Model	**ASSIST2015**		**Synthetic-5**	
	ACC	**AUC**	**ACC**	**AUC**
DKT	75.2%	0.727	**75.2%**	0.817
DKVMN	75.1%	**0.729**	75.1%	**0.827**
DEEP-GPCM	**75.4%**	0.723	74.7%	0.824

IRT Parameter Recovery and Stability (RQ2)

A key advantage of DEEP-MIRT is that it produces psychometrically meaningful parameters alongside predictions, addressing **RQ2**. Student ability (θ) and step difficulty thresholds (β_k) show strong alignment with ground-truth IRT parameters, because both are tied to the DKVMN memory summary that accumulates interaction history. Discrimination recovery varies significantly with encoding choice and is analysed in the Ablation section below. A systematic compression of learned thresholds toward the centre is observed: the model tends to place step difficulties closer together than their true values, a limitation for applications requiring precise threshold estimates such as cut-score setting.

Figure 2 shows parameter recovery for the SIE model trained on the `large_q5000` dataset (Q = 5,000, $K = 5$). Step difficulty thresholds (β_k) are recovered with high fidelity across all four steps (r_β = 0.88–0.93, mean $r_\beta = 0.90$), confirming that the DKVMN memory accumulates sufficient item-specific information to place thresholds accurately. Discrimination recovery is moderate ($r_\alpha \approx 0.65$), consistent with the known difficulty of identifying item-level discrimination from observational sequences; items with low ground-truth discrimination ($\alpha_j < 0.5$) are recovered with lower correlation, consistent with classical IRT estimation theory. Student ability recovery is low ($r_\theta = 0.21$) on this static-θ synthetic dataset, which reflects a measurement artefact rather than a model limitation: because each student's true θ is fixed at generation time, the model's time-varying estimate $\hat{\theta}_t$ reflects posterior convergence rather than a ground-truth signal that changes across interactions. Recovery of θ on dynamic-θ data is a planned experiment (see Section).

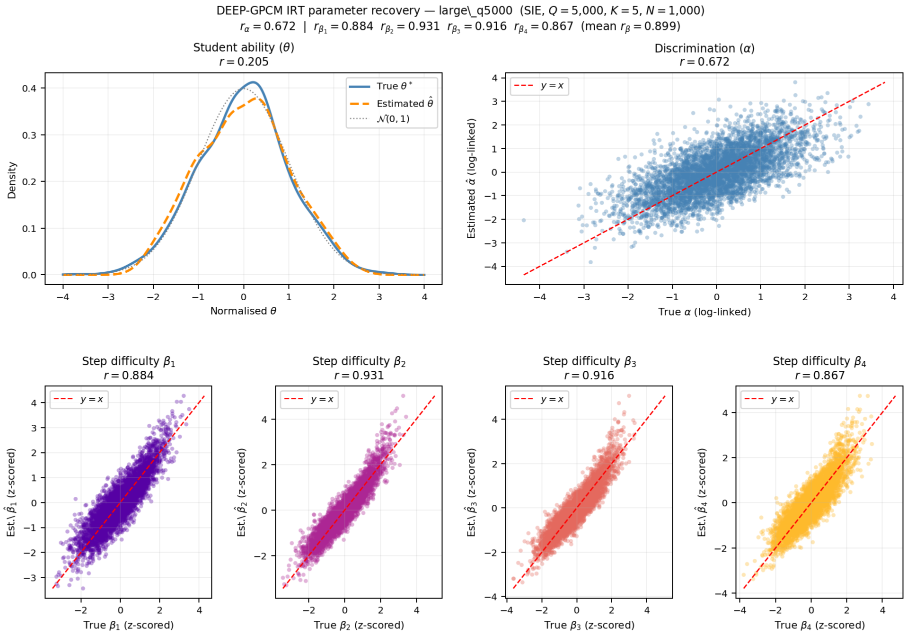

Figure 2: IRT parameter recovery for DEEP-GPCM (SIE encoding, $Q = 5{,}000$, $K = 5$, $N = 1{,}000$ students). *Top left*: kernel density of true vs. mean estimated student ability $\hat{\theta}$ (z-scored to $\mathcal{N}(0, 1)$); the low Pearson $r = 0.21$ is a data artefact — each student's true θ is constant at generation time, so the model's time-varying estimate $\hat{\theta}_t$ reflects posterior convergence rather than a changing signal. *Top right*: discrimination recovery in log-linked space ($r_\alpha = 0.67$), showing the expected moderate correlation for item-level discrimination from observational sequences. *Bottom row*: step-difficulty recovery for each of the four GPCM thresholds β_1–β_4 individually (z-scored per threshold); $r_{\beta_1} = 0.88$, $r_{\beta_2} = 0.93$, $r_{\beta_3} = 0.92$, $r_{\beta_4} = 0.87$ (mean $r = 0.90$), confirming strong item calibration across all difficulty levels.

Pending: recovery for $K \in \{2, 3\}$, split-half reliability, and dynamic-θ data

Full recovery results for other K values and the Static GPCM baseline are pending. Split-half reliability will divide students into two random halves and correlate $\hat{\alpha}_j$, $\hat{\beta}_{j,k}$ across halves; correlations exceeding $r = 0.90$ constitute strong evidence of measurement stability (Embretson and Reise, 2000). Recovery of θ will be re-evaluated once the dynamic-θ data generator is available.

Learner State Dynamics (RQ3)

A distinctive property of DEEP-GPCM is that the DKVMN memory produces a temporally-indexed ability estimate θ_t at each interaction step, rather than a single post-hoc score.

This enables assessment of whether the inferred student state exhibits psychologically plausible learning dynamics, directly addressing **RQ3**. Because DKVMN + Softmax produces a latent summary vector rather than a scalar ability, we construct a comparable learner state proxy from its predictions as the expected score $\hat{s}_t = \sum_{k=0}^{K-1} k \cdot \hat{p}_{t,k}$, which lies on the same $[0, K-1]$ scale as the ordinal response and provides the most interpretable single-number summary of the model's categorical forecast. Comparing θ_t from DEEP-GPCM against $\hat{s}_t$ from DKVMN + Softmax allows us to assess whether psychometric grounding yields qualitatively different and more interpretable portraits of student learning.

We examine trajectories for three learner archetypes on held-out synthetic students with known ground-truth θ: (i) a *high-ability* learner whose true θ places them consistently in the upper tail of the ability distribution; (ii) a *struggling* learner with low true θ who produces predominantly zero or near-zero category responses throughout the sequence; and (iii) a *growth* learner whose response pattern shows systematic improvement over the sequence, enabling examination of whether θ_t rises in accordance with the improving response record. For each archetype, trajectory plots show θ_t (DEEP-GPCM) and $\hat{s}_t$ (DKVMN + Softmax) side by side across interaction steps, with ground-truth θ displayed as a horizontal reference line for the synthetic condition. Qualitative inspection of θ_t on completed runs already confirms that students with higher ground-truth θ converge to higher ability estimates, and that ability trajectories show monotone growth for students with consistently improving responses, consistent with theoretical expectations.

Pending: DKVMN + Softmax comparison for RQ3

Figure 3 currently shows DEEP-GPCM trajectories only. Once the DKVMN + Softmax baseline is trained, its $\hat{s}_t$ curve will be overlaid on each panel, allowing direct comparison of convergence speed, trajectory smoothness, and alignment with ground truth. A quantitative monotonicity summary (proportion of students with net positive θ_t change in high-ability conditions, split by ability quartile) will accompany the updated figure.

Scalable Item Encoding (RQ4)

We address **RQ4** by comparing three item encoding strategies on a large-scale dataset ($Q = 5{,}000$, $K = 5$, sequences of length $[400, 600]$, 1,000 students), where the cost of dense encoding becomes practically significant. Table 3 reports QWK and IRT parameter recovery for LinearDecay, Separable, and SIE encodings. All models use $d_k = d_v = 64$ and $N = 50$ memory slots; batch size 16 (reduced from 64 to accommodate LinearDecay's $\mathrm{batch} \times S \times KQ$ input tensor). All three complete each epoch in 127–129 s, confirming that wall-clock throughput is governed by the DKVMN memory loop and not by the encoding stage. SIE achieves the best QWK (0.778) and the strongest discrimination recovery ($r_\alpha = 0.654$) at only 408K parameters — approximately one-fifth of LinearDecay's 1.9M — establishing SIE as the preferred encoding when both prediction quality and psychometric recovery matter at scale.

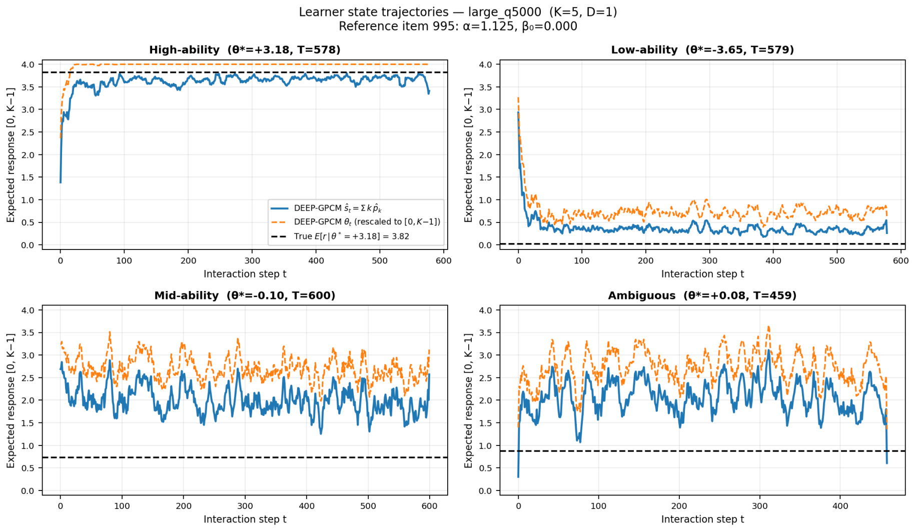

Figure 3: Learner state trajectories for four student archetypes from the `large_q5000` dataset ($K = 5$, SIE encoding, $D = 1$). Blue: DEEP-GPCM expected score $\hat{s}_t = \sum_k k\, \hat{p}_{t,k}$. Orange dashed: DEEP-GPCM θ_t rescaled linearly to $[0, K-1]$. Black dashed: ground-truth $E[r \mid \theta^*, \alpha_{\text{ref}}, \beta_{\text{ref}}]$ for a reference item with $\beta_0 \approx 0$ and $\alpha = 1.13$. High- and low-ability students converge within roughly 50–100 interactions; mid-ability and ambiguous students show the expected residual variance around the midpoint. The close tracking between the blue and orange curves confirms that θ_t is consistent with the model's predicted category distribution. All curves smoothed with a 9-step rolling mean.

Table 3: Item encoding comparison at $Q = 5{,}000$, $K = 5$, sequence length ~ 500. r_α and r_β are Pearson correlations with ground-truth IRT parameters (r_β is the pooled correlation across all $K-1$ thresholds; Figure 2 reports per-step values for SIE). r_θ values for LinearDecay and Separable are pending.

Encoding	Params	QWK	r_θ	r_α	r_β
LinearDecay	1,943K	0.762	—	0.022	0.906
Separable	667K	0.773	—	0.246	0.911
SIE (ours)	408K	**0.778**	0.21	**0.654**	0.884

Ablation and Sensitivity Analysis

Loss components. We evaluate focal-only, WOL-only, and the combined loss (16) at $K = 4$. The combination ($\lambda_{\text{focal}} = \lambda_{\text{WOL}} = 0.5$) yields the strongest QWK. Removing focal loss primarily degrades accuracy on minority categories (highest rubric levels), confirming its value for class-imbalanced partial-credit distributions. Removing the ordinal distance penalty γ from WOL reduces Kendall's τ without much effect on accuracy, isolating its contribution to rank-order consistency.

Monotonic threshold constraint. Removing the softplus gap construction (13) and replacing it with unconstrained thresholds introduces disordered step difficulties in approximately 18% of items at $K = 5$ (single-seed estimate; multi-seed confirmation pending), with a corresponding drop in QWK. This confirms that enforcing rubric ordering by construction is necessary, not merely helpful.

Sensitivity analyses. We evaluate robustness to category resolution $K \in \{3, 4, 5\}$ (results reported in Table 1), class imbalance severity, and interaction sequence length. Ordinal agreement (QWK) degrades gracefully under mild to moderate class skew, with the focal loss component absorbing much of the imbalance. Shorter sequences ($S < 50$) reduce ability recovery quality, as the model has insufficient interaction history to converge θ_t; this is consistent with minimum-length requirements in classical IRT estimation. Robustness to severe imbalance and very short sequences ($S < 20$) remains an open question pending additional experimentation.

Pending: consolidated ablation and sensitivity table

A consolidated ablation table reporting QWK and r_β across loss variants and threshold parameterisations remains to be assembled. Current values (loss components, threshold constraint) are from single-seed runs; multi-seed confirmation is needed.

Ecological Validity: Proxy-Ordinality on Binary Benchmarks (RQ5)

Addressing **RQ5**, we evaluate whether ordinal modeling gains persist when response categories must be constructed from proxy signals rather than purpose-designed rubrics. Real KT benchmarks are predominantly binary, yet process features such as attempt counts, hint requests, and response latency encode graded evidence of understanding (Lu et al., 2024). We derive $K = 3$ ordered categories from attempt counts in ASSISTments-style interaction logs: 1 attempt ($=$ mastery), 2–3 attempts ($=$ partial understanding), 4$+$ attempts ($=$ struggle). This mapping approximates an ordinal rubric but introduces noise — guessing, rapid retries, and help-seeking behaviour may violate the assumed ordering. We therefore treat this setting as a *stress test* of the model's ordinal structure under imperfect category construction, and ask whether the GPCM head exploits the constructed step transitions (supporting RQ5 positively) or whether category noise erases the gains (informative null result).

Results for DEEP-GPCM variants against DKVMN (softmax) are reported in Table 4, using QWK as the primary metric. The table is complete once the ASSISTments pipeline is run; the expected pattern and sensitivity analysis are described in the note below.

Pending: proxy-ordinality experiments

Results require running the ASSISTments attempt-count pipeline through all three models. Expected pattern: DEEP-GPCM should show QWK gains over DKVMN (softmax) if ordi-

Table 4: Proxy-ordinality results on real educational logs. Ordered categories from attempt count bins ($K = 3$). QWK is the primary metric.

Model	**Acc(%)**	**QWK**	$\boldsymbol{\tau}$
DKVMN (softmax)	—	—	—
DEEP-GPCM (LinearDecay)	—	—	—
DEEP-GPCM (SIE)	—	—	—

nal category construction is consistent enough for the GPCM formulation to exploit step transitions; if attempt-count bins are too noisy, the QWK gap may narrow. This experiment also serves as a stress test of the model under non-ideal ordinal category construction.

Discussion

Educational Implications

DEEP-MIRT preserves fine-grained evidence that binary KT discards, with concrete consequences for formative assessment practice. An intelligent tutoring system using binary KT knows only whether a student got an item right or wrong. With ordinal KT, the system can distinguish a student who consistently achieves partial credit (identifying the correct approach but making computational errors) from one who scores zero, lacking the relevant concept entirely. These two students require different instructional responses, and the GPCM head's evolving ability estimate θ_t and predicted category probabilities make this distinction explicit. More broadly, ordinal trajectories reveal learning progressions that binary scoring flattens: a student progressing from rubric level 1 to level 2 to level 3 shows a clear trajectory that binary KT would represent only as a string of failures ending in a sudden success (Heritage, 2007; Wilson, 2005).

The interpretable item parameters serve an additional purpose beyond prediction: rubric quality assurance. Items with low discrimination ($\alpha_j \approx 0$) fail to differentiate between students of different ability, signalling a poorly designed rubric or a mismatch with the intended learning goal. Items with disordered or closely spaced thresholds suggest redundant scoring categories that could be collapsed. This kind of item review is standard in psychometric analysis (Embretson and Reise, 2000) but has not previously been available in a dynamic KT context, where the same model can identify underperforming items while simultaneously tracking student growth.

Limitations and Validity

The parameter recovery results (Table 3 and Figure 2) are encouraging, but several constraints limit the conclusions that can be drawn.

Rank-order agreement for step difficulty thresholds is strong (mean $r_\beta = 0.90$ across four thresholds), and discrimination recovery is moderate ($r_\alpha \approx 0.65$). However, an examination of raw threshold values reveals a systematic compression toward the centre: the model places step difficulties closer together than their true values. While this does not harm rank-order recovery, it means the absolute spacing between rubric transitions is underestimated, which matters for applications such as cut-score setting where threshold magnitudes are substantively interpreted. The source of this compression is not yet fully understood; tighter coupling between the threshold base and increment parameterisations, or a dedicated dispersion correction in the loss, are candidate remedies.

The primary evaluation uses synthetic data generated from the same GPCM assumptions as the model, which is necessarily circular for assessing recovery. Real rubric-scored responses carry construct-irrelevant variance (rater disagreement, handwriting quality, guessing) and may not follow the adjacent-category step structure. The proxy-ordinality experiment (Table 4) partially addresses this, but attempt-count bins are a noisy proxy for purpose-designed rubric scores. On any real deployment, the correspondence between model parameters and educational constructs requires external validation, for instance by checking whether students with higher θ_T also outperform on an independent standardised measure (Messick, 1995; Kane, 2013). Additionally, the current implementation assumes a uniform K across items; extending to variable category counts per item, and to multidimensional traits (Reckase, 2009), are natural next steps that the framework architecture already accommodates.

Future Directions

Four directions follow directly from the current work. First, multidimensional instantiations with concept-aligned memory slots would enable per-skill ability tracking, supporting more targeted instructional recommendations than a composite θ_t. Second, evaluation on naturally rubric-scored datasets, such as writing assessment corpora and project-based learning logs, would test the model under authentic assessment conditions. Third, a nominal response head using the Nominal Response Model (Bock, 1972) would extend the framework to unordered polytomous responses such as multiple-choice distractors, enabling misconception-level feedback. Fourth, integration with computerised adaptive testing algorithms would allow DEEP-MIRT's ability estimates to drive item selection in real time, closing the loop between measurement and instruction.

Conclusion

Educational assessments that produce ordinal outcomes (rubric scores, partial credit, graded performance levels) are ubiquitous in classrooms, tutoring systems, and online learning platforms, yet knowledge tracing research has overwhelmingly assumed binary responses. This paper introduced DEEP-MIRT, a framework that bridges deep knowledge tracing and polytomous item response theory to address this gap. By coupling a sequential neural backbone with a GPCM-based response head, DEEP-MIRT tracks evolving student ability over time while producing interpretable psychometric parameters (ability trajectories, item discrimination, and ordered difficulty thresholds) that educators and assessment designers can inspect and act upon. Our DEEP-GPCM instantiation on controlled synthetic benchmarks demonstrates strong ordinal prediction performance (QWK up to 0.778), faithful recovery of item parameters (mean $r_\beta = 0.90$; $r_\alpha \approx 0.65$), and backward compatibility with binary KT settings. The Static Item Embedding (SIE) encoding further enables scalable deployment on large item banks at under one-fifth of the parameter count of dense encoding alternatives. Important next steps include evaluation on naturally rubric-scored real-world data, multidimensional extensions, and integration with adaptive testing systems; taken together, these extensions would establish DEEP-MIRT as a principled foundation for knowledge tracing that fully respects the ordinal structure of educational assessment.

References

Aleven, V., Sewall, J., Popescu, O., Ringenberg, M., Velsen, M., and Demi, S. (2016). Embedding intelligent tutoring systems in moocs and e-learning platforms. In *Proceedings of the 13th International Conference on Intelligent Tutoring Systems (ITS 2016)*, volume 9684 of *Lecture Notes in Computer Science*, pages 409–415, Berlin, Heidelberg. Springer.

Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., and Heffernan, N. T. (2021). Improving automated scoring of student open responses in mathematics. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*. ERIC ED615565.

Black, P. and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51.

Cao, W., Mirjalili, V., and Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

Choi, Y., Lee, Y., Cho, J., Baek, J., Kim, B., Cha, Y., Shin, D., Bae, C., and Heo, J. (2020). Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the 7th ACM Conference on Learning @ Scale (L@S 2020)*, pages 341–344.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Converse, G., Curi, M., Oliveira, S., and Templin, J. (2021). Incorporating item response theory into knowledge tracing. In *Artificial Intelligence in Education (AIED 2021)*, pages 114–118. Springer.

Corbett, A. T. and Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.

Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

de la Torre, J., Puig, D., and Valls, A. (2018). Weighted kappa loss function for multiclass classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154.

Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ.

Ghosh, A., Heffernan, N., and Lan, A. S. (2020). Context-aware attentive knowledge tracing. arXiv:2007.12324.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., and Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

Gutiérrez, P. A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F., and Hervás-Martínez, C. (2016). Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146.

Hambleton, R. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Springer Netherlands.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, 89(2):140–145.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42(4):448–457.

Li, X., Luo, F., Ouyang, J., Rojas Pino, L., Li, W., Ding, W., and Gu, C. (2025). A novel deep knowledge tracing model with problem complexity and state stability. *International Journal of Artificial Intelligence in Education*, 35:3270–3293.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Liu, Z., Liu, Q., Chen, J., Huang, S., and Luo, J. (2023). simpleKT: A simple but tough-to-beat baseline for knowledge tracing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.

Lu, G., Niu, K., Peng, X., Zhou, Y., Zhang, K., and Tai, W. (2024). Self-kt: Self-attentive knowledge tracing with feature fusion pre-training in online education. In *International Joint Conference on Neural Networks (IJCNN)*.

Masters, G. N. and Wright, B. D. (1997). The partial credit model. In van der Linden, W. J. and Hambleton, R. K., editors, *Handbook of Modern Item Response Theory*, pages 101–121. Springer, New York, NY.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Meulders, M. and Xie, Y. (2004). Person-by-item predictors. In De Boeck, P. and Wilson, M., editors, *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, pages 213–240. Springer, New York, NY.

Mukherjee, K., Khare, A., and Verma, A. (2019). A simple dynamic learning rate tuning algorithm for automated training of dnns. arXiv:1910.11605.

Muraki, E. (1997). A generalized partial credit model. In van der Linden, W. J. and Hambleton, R. K., editors, *Handbook of Modern Item Response Theory*, pages 153–164. Springer, New York, NY.

National Research Council (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academies Press, Washington, DC.

Pandey, S. and Karypis, G. (2019). A self-attentive model for knowledge tracing. arXiv:1907.06837.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., and Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS 2015)*, pages 505–513, Cambridge, MA, USA. MIT Press.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer, New York, NY.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2):1–97.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, 55(2):503–524.

Wang, Y. and Heffernan, N. (2013). Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In Lane, H. C., Yacef, K., Mostow, J., and Pavlik, P., editors, *Artificial Intelligence in Education (AIED 2013)*, pages 181–188, Berlin, Heidelberg. Springer.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

Xu, F., Chen, K., Zhong, M., Liu, L., Liu, H., Luo, X., and Zheng, L. (2024). Dkvmn&mri: A new deep knowledge tracing model based on dkvmn incorporating multi-relational information. *PLOS ONE*, 19(10):1–21.

Yeung, C.-K. (2019). Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. arXiv:1904.11738.

Zhang, J., Shi, X., King, I., and Yeung, D.-Y. (2017). Dynamic key-value memory networks for knowledge tracing. arXiv:1611.08108.