Article

Beyond Black-Box Deep Knowledge Tracing: Transformers with Representational Grounding for Pedagogical Interpretability

Concha Labra [1]*, Olga C. Santos [2]

[1] Universidad Nacional de Educación a Distancia: Madrid, Madrid, ES; clabra@dia.uned.es
[2] Universidad Nacional de Educación a Distancia: Madrid, Madrid, ES; ocsantos@dia.uned.es
* Author to whom correspondence should be addressed.

Abstract

This study introduces iDKT, an interpretable-by-design Transformer model that utilizes *Representational Grounding* to align deep latent representations with educational constructs, leveraging the high accuracy of deep knowledge tracing models while addressing their inherent lack of interpretability. We introduce a formal validation framework to verify the alignment of iDKT's internal representations and, using Bayesian Knowledge Tracing (BKT) as a reference, evaluate the model across multiple educational datasets. Results demonstrate that iDKT maintains state-of-the-art predictive performance while yielding additional interpretable insights at a significantly higher granularity than those provided by the reference model. Specifically, iDKT identifies student-level initial knowledge and learning velocities, providing mastery estimates that are more sensitive to the nuances of individual behavioral patterns than those produced by standard BKT. These individualized insights enable precise diagnostic placement and dynamic pacing, allowing adaptive learning environments to tailor instruction to each student's unique learning profile with enhanced precision. This work offers both a robust methodology for evaluating the interpretability of Transformer-based models and a practical tool for improving educational effectiveness through data-driven personalization.

Keywords: deep knowledge tracing; transformer; interpretability; Bayesian Knowledge Tracing; educational data analysis; personalized learning

1. Introduction

Knowledge Tracing [1] is a fundamental task in the fields of Artificial Intelligence in Education, Intelligent Tutoring Systems and Massive Open Online Courses. Its primary objective is to model a student's dynamic knowledge state over time based on their history of interactions with learning materials, enabling systems to predict future performance and provide personalized instruction. As educational environments become increasingly diverse and digital, the ability to accurately track and interpret student mastery has become a critical requirement for scalable, effective education.

Historically, the field has been dominated by two distinct paradigms. The first, exemplified by Bayesian Knowledge Tracing (BKT) and its variants [2], relies on probabilistic graphical models that explicitly represent knowledge states. BKT models are intrinsically interpretable, being based on parameters such as initial knowledge, learning rate, or slipping and guessing probabilities that map directly to pedagogical constructs, allowing educators to understand how they work and trust their decisions. However, this interpretability

Received:
Revised:
Accepted:
Published:

 https://doi.org/10.3390/1010000

comes at the cost of a simplicity that often limits its predictive power, making them struggle to capture the complex, non-linear dependencies often present in educational datasets.

The second paradigm emerged with the advent of Deep Knowledge Tracing (DKT) [3], which uses different variants of deep learning techniques from the initial Recurrent Neural Networks to current Transformers [4] to model student interactions. These models have achieved state-of-the-art predictive performance, significantly outperforming classical approaches by leveraging the high capacity of deep learning models that allows them to learn complex patterns [5]. Yet, this predictive power has come at a significant cost: interpretability. Deep learning models are notoriously opaque "black boxes," where the learned representations are distributed across high-dimensional latent spaces that bear no direct correspondence to constructs with a clear semantic meaning. This lack of transparency creates a trust gap for practitioners, who cannot easily discern why a model predicts a student has failed or succeeded, nor can derive actionable pedagogical insights from the model's internal state [6].

Current efforts to bridge this gap typically rely on post-hoc explainability methods, such as weights visualization or perturbation analysis [7,8]. While valuable for debugging, these techniques often provide only a superficial view of the model's decision-making process and do not guarantee that the learned representations align with meaningful constructs. Moreover, their application and interpretation require technical deep learning expertise, limiting their accessibility to practitioners without this specialized knowledge.

To address these limitations, we propose a shift towards interpretability-by-design, inspired by the emerging paradigm of Theory-Guided Data Science (TGDS) [9]. In TGDS, maintaining consistency with theoretical postulates is an architectural constraint rather than an afterthought. By integrating extensive domain knowledge, TGDS-based models can be constrained to learn representations that are both theoretically plausible and highly predictive. While this approach has been applied mostly to science—and specifically to physics [10]—we adapt it here to the educational domain.

Standard TGDS implementations typically rely on auxiliary loss functions to incorporate formal knowledge expressed as rules, algebraic constraints, or differential equations [11]. We propose a novel approach called *Representational Grounding* that, in contrast, utilizes auxiliary losses operating on projections of the Transformer's embeddings. This mechanism enables the model to learn representations that are consistent with semantically meaningful constructs.

The major contributions of this work are as follows:

- Proposal of Representational Grounding, a novel method that overcomes the black-box nature of Transformers by providing interpretability-by-design.
- Introduction of a formal validation framework to quantify interpretability via representational alignment, enabling a systematic characterization of the trade-off between reference fidelity and predictive performance.
- Application of Representational Grounding to the development of iDKT, a new type of knowledge tracing models that leverage the high accuracy inherent in deep learning while achieving pedagogical interpretability.
- Empirical demonstration of iDKT benefits by showing how it captures granular, student-specific insights—such as individualized initial knowledge and learning rates—that are beyond the capabilities of simpler models such as BKT.

The remainder of this paper is structured as follows. Section 2 reviews related work on knowledge tracing, deep learning interpretability, and theory-guided data science. Section 3 describes the proposed iDKT architecture and the *Representational Grounding* framework. Sections 4 and 5 detail the individualized embedding mechanism and the multi-objective loss system, respectively. Section 6 presents the results of the experimental validation, interpretability results, and the analysis of individualization granularity. Finally, Section 7 concludes the paper.

2. Related Work

2.1. Deep Knowledge Tracing

2.2. Bayesian Knowledge Tracing

2.2.1. Parameters of the Standard Model

2.2.2. Individualization

2.3. Deep Learning Interpretability

2.4. Theory-Guided Data Science

3. iDKT Model

We propose an Interpretable Deep Knowledge Tracing (iDKT) model with a Transformer-based architecture designed to bridge the gap between the high predictive capacity of deep learning and the intrinsic interpretability of simpler models such as Bayesian Knowledge Tracing (BKT). Unlike standard black-box models, iDKT utilizes a novel mechanism called Representational Grounding to anchor its latent representations to the conceptual space of a interpretable model choosen as reference.

The core architecture of iDKT, as illustrated in Figure 1, extends the standard Transformer framework [4] by incorporating specialized components for Representational Grounding, primarily integrated within the embedding layers and the multi-objective loss pipeline.

1. **Input Data**: This stage handles the ingestion of student interactions (concepts, questions, and binary responses $r \in \{0,1\}$). Crucially, it also loads values from the BKT reference model, including performance predictions (p_{BKT}) and per-skill parameters such as initial mastery (l_0) and learning transition rates (T), which serve as grounding targets for the model.
2. **Embeddings**: Task and history information are projected into a continuous embedding space. We then apply the individualization transformations described in Section 4 where base embeddings are augmented by student-specific parameters: a Proficiency Offset (lc) and a Rate Augmentation (tc). This produces the Transition Gap (individualized task representation x') and the Transition Gain (individualized interaction history y'), effectively defining the difficulty of a task relative to the student's baseline proficiency.
3. **Transformer Core**: The model employs a dual-encoder architecture followed by a cross-attention decoder. One encoder processing the individualized interaction history ($y'_{1:t-1}$) to capture global student behavior, while a task encoder processes the current individualized task (x'_t). The decoder uses multi-head cross-attention to synthesize these streams into a latent Proficiency Context, representing the student's current specialized knowledge state for the target task.
4. **Output Stage**: The latent Proficiency Context is passed through a multi-layer perceptron (MLP) that maps the deep representations to the final output space. The final output is the iDKT prediction (p_{iDKT}), representing the probability that the student will respond correctly to the current task.
5. **Loss Functions (Grounding Pipeline)**: During training, the architecture is supervised not only by the prediction loss (L_{sup}) against ground truth outcomes but also by an alignment pipeline. This includes the Alignment Loss (L_{ref}) which penalizes deviations from the BKT prediction, and Parameter Losses (L_{init}, L_{rate}) that ground the model's internal proficiency and rate parameters to their BKT-derived theoretical counterparts.

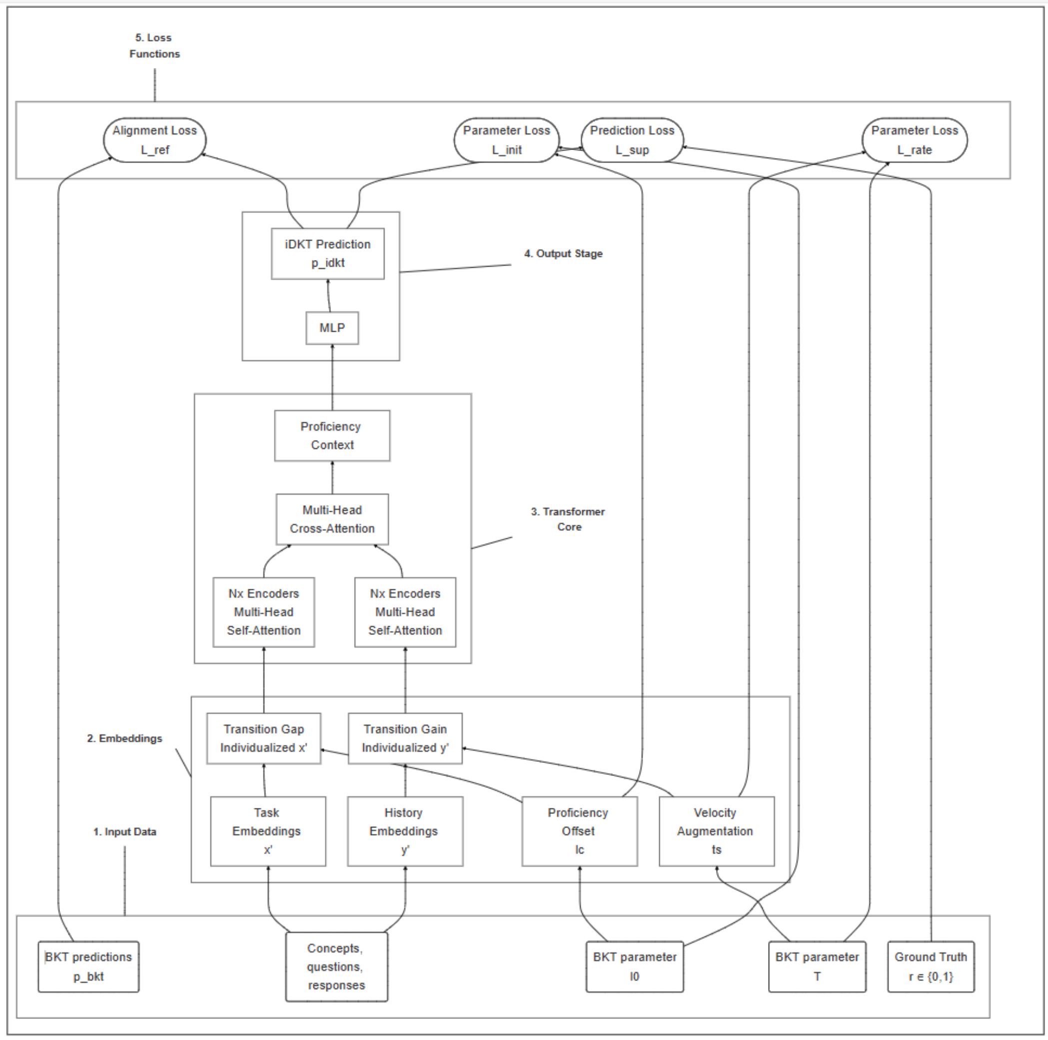


Figure 1. The iDKT Architecture. The diagram illustrates the five functional stages: (1) Input Data ingestion including BKT targets, (2) Individualized Embeddings, (3) the Transformer Core, (4) the MLP-based Output, and (5) the Loss Functions for representational grounding.

4. Embeddings

In standard educational datasets, such as ASSISTments 2009, ASSISTments 2015, Algebra 2005, and others [12], student interactions are recorded at the level of specific questions or tasks, each of which is associated with one or more underlying concepts or knowledge components. This structure reflects the fact that proficiency in a concept (e.g., the Pythagorean Theorem) is acquired through interactions with a diverse range of tasks. While all tasks involving a concept share the same semantic core, they differ in their specific manifestations—most notably in their intrinsic difficulty or complexity. Therefore, an effective representation must capture both the shared identity of the concept and the unique deviation of the specific task.

In Transformer-based models [4] we can operationalize this principle representing the tasks as embedding vectors:

$$x' = c + u \cdot d \quad \text{(Task)} \tag{1}$$

In this formulation, c acts as the *Concept Anchor*, a vector representing the invariant semantic identity of the concept while the vector d represents the learnable *Question Variation Axis*, defining the specific direction of the "transition gap". The scalar u serves as the *Relational Magnitude*, representing the question's specific relative difficulty compared to other questions involving the same concept. Consequently, rather than encode arbitrary embeddings for every question, we encode the vector sum of two distinct components with clear semantic meaning: a base *concept identity* and a *difficulty shift*.

In a similar way, we can operationalize the interactions between questions and students as embedding vectors:

$$y' = e + u \cdot (f + d) \quad \text{(Interaction History)} \tag{2}$$

Here e represents the *Interaction Base*, which is a combined representation of concept c and the binary outcome r (correct/incorrect), while f represents the *Interaction Variation Axis*, which is similar to the Question Variation Axis (d_c) but is specific to the interaction between a question and a student. The inclusion of d_c in the interaction shift ensures that the difficulty vectors are consistent across both questions (x') and interactions (y').

Extending this rationale, we can enrich the x' embeddings by integrating additional components with explicit semantic significance. Specifically, by adopting Bayesian Knowledge Tracing (BKT) as a reference model, we can incorporate vectors corresponding to its core theoretical parameters—Initial Knowledge (L_0) and Learning Rate (T)—thereby grounding the deep representation in established pedagogical constructs.

To get individualized values for these parameters, we decompose them into population-level bases and student-specific deviations:

$$l_c = L_0 + k_s \cdot d_k \quad \text{(Personalized Initial Knowledge)} \tag{3}$$

$$t_c = T + v_s \cdot d_v \quad \text{(Personalized Learning Rate)} \tag{4}$$

where l_c is the personalized initial knowledge for concept, t_c is the personalized learning rate for concept, L_0 and T are the population-level base embeddings, d_k and d_v are the learnable variation axes vectors (similar to the difficulty axis d), and k_s and v_s are the scalar student-specific deviations learned for each individual.

We include these vectors to get the final input embedding for the encoder and decoder components of the Transformer:

$$x' = (c + u \cdot d) - l_c \quad \text{(Individualized Task)} \tag{5}$$

$$y' = (e + u \cdot (f + d)) + t_c \quad \text{(Individualized Interaction History)} \tag{6}$$

where c represents the concept embedding, u the question-specific difficulty shift, d the task variation axis, e the interaction base, f the interaction variation axis, l_c the personalized initial knowledge, and t_c the personalized learning rate.

The rationale for using difference for the individualized task (x') and sum for the interaction history (y') is due to their distinct semantic roles:

- x' represents the *Transition Gap*: *Difficulty* $-$ *Proficiency*. Under this relational logic, objective task difficulty is offset by prior proficiency, ensuring that task demands are defined relative to the subject's baseline. This formulation captures the difference between the task requirements and the current state, representing the residual gap after accounting for latent proficiency.
- y' represents the *Transition Gain*: *Interaction* $+$ *Rate*. The history encoder accumulates evidence from interactions, applying a consistent relational logic where the signal value is augmented by latent rate, ensuring that interaction outcomes are defined relative to the subject's pace. Under this formulation, the total value encompasses not only the interaction outcome but also the rate of progress through the state trajectory, as this incremental gain serves as a robust indicator of future performance. This approach, therefore, captures latent progression by augmenting observed outcomes with transition rate, thereby reflecting individualized acquisition rates.

5. Loss Functions

The model is trained using a multi-objective loss function designed to ensure that the high-capacity Transformer remains aligned with pedagogical principles through Representational Grounding. The total loss $\mathcal{L}_{total}$ is defined as a weighted sum of different loss components described in detail below.

$$\mathcal{L}_{total} = L_{sup} + \lambda_{ref} L_{ref} + \lambda_{init} L_{init} + \lambda_{rate} L_{rate} + L_{reg} \tag{7}$$

5.1. Supervised Alignment (L_{sup})

The primary objective L_{sup} uses standard Binary Cross-Entropy (BCE) between the iDKT performance predictions $\hat{y}_t$ and the observed ground truth outcomes $r_t \in \{0, 1\}$. This loss ensures predictive accuracy by minimizing the deviance from observed student behavior:

$$L_{sup} = -\frac{1}{N} \sum_{t=1}^{N} [r_t \log(\hat{y}_t) + (1 - r_t) \log(1 - \hat{y}_t)] \tag{8}$$

5.2. Representational Grounding

The grounding losses ($L_{ref}, L_{init}, L_{rate}$) use Mean Squared Error (MSE) to anchor deep representations to BKT-derived values. Specifically, L_{ref} forces behavioral predictions to stay close to the theoretical baseline, while L_{init} and L_{rate} ground the individualized parameters l_c and t_c in meaningful educational starting points and acquisition paces, respectively. Instead of arbitrary latent weights, the model's internal states are projected through a sigmoid activation $\sigma(\cdot)$ and compared directly to the reference values:

$$L_{ref} = \text{MSE}(\hat{y}, p_{BKT}) \tag{9}$$

$$L_{init} = \text{MSE}(\sigma(\bar{l}_c), L0_{BKT}) \tag{10}$$

$$L_{rate} = \text{MSE}(\sigma(\bar{t}_c), T_{BKT}) \tag{11}$$

where $\bar{l}_c$ and $\bar{t}_c$ are the average across the feature dimension of the individualized embeddings for proficiency and rate, respectively. This formulation forces the deep representation to be not only predictive but also semantically consistent with the reference constructs.

5.3. Inductive Bias Regularization (L_{reg})

While the grounding losses anchor the global position of the latent space to the BKT parameter estimations, L_{reg} ensures that student-level individualization is *parsimonious*. This loss acts directly on the individualization parameters (u_q, k_s, v_s) to ensure that the model only deviates from the theoretical prior when functionally necessary. We apply distinct L_2 penalties to the scalar parameters governing variation:

$$L_{reg} = \lambda_{\text{u}} \sum_{q \in Q} u_q^2 + \lambda_{\text{k}} \sum_{s \in S} k_s^2 + \lambda_{\text{v}} \sum_{s \in S} v_s^2 \tag{12}$$

where u_q represents item difficulty, k_s is the student-specific knowledge gap, and v_s is the learning rate deviation. This formulation implements a *normal student prior*: the model assumes every subject adheres to the population-level parameters derived from the BKT reference unless their unique interaction history provides sufficient signal to justify the regularization cost.

6. Results and Discussion

6.1. Experimental Setup

We implemented the iDKT model in `PyTorch` and used the benchmark library `PYKT` [12] to leverage standardized data preprocessing, dataset splitting and benchmarking of baseline models. For training, we used standard 5-fold cross-validation with an 80/20 train/test split. The model was trained using the Adam optimizer with a learning rate of $1e-4$, a batch size of 64, and a dropout rate of 0.2 to prevent overfitting. The maximum number of epochs was set to 200, with an early stopping mechanism (patience=10) to terminate training if validation performance plateaued.

The Transformer architecture was configured with an embedding dimension d_{model} of 256, 8 attention heads, and 4 encoder/decoder blocks. The feed-forward dimension d_{ff} was set to 512. Regularization penalties for individualization parameters (L_2 on u_q, k_s, v_s) were all set to $1e-5$. For reproducibility, all experiments were seeded (seed=42) and executed on NVIDIA A100 GPU infrastructure. Predictive performance was evaluated using Area Under the ROC Curve (AUC) and Accuracy (ACC), while interpretability was assessed using the metrics defined in Section 6.5.

6.2. Datasets

We did the evaluation of iDKT with these 5 widely used datasets:

• ASSISTments2009: A dataset consisting of math exercises, collected from the free online tutoring ASSISTments platform in 2009-2010. It is one of the most widely used and has been the standard benchmark for many years [13].

• ASSISTments2015: This dataset was collected from the ASSISTments platform in the year of 2015. It has the largest number of students among the other ASSISTments datasets [13].

• Algebra2005: A dataset from the KDD Cup 2010 EDM Challenge containing questions from the Carnegie Learning Algebra system deployed 2005-2006 [14].

• Bridge2006: A dataset from the KDD Cup 2010 EDM Challenge with the Carnegie Learning Bridge to Algebra system, deployed 2006-2007 [15].

• NIPS34: A dataset collected from the Eedi platform [16] containing answers to multiple-choice diagnostic math questions for the Tasks 3 & 4 at the NeurIPS 2020 Education Challenge.

6.3. Predictive Performance

To evaluate the predictive performance of the iDKT model, we used the Area Under the Curve (AUC) and the Classification Accuracy (ACC) of the models on the test set with the 5 datasets described in Section 6.2. The results are shown in the Table 1.

Table 1. Predictive Performance of the iDKT model across 5 datasets.

Model	AS2009	AS2015	Algebra2005	Bridge2006	NIPS34
iDKT	0.8358	0.7252	0.9281	0.8092	0.7987

When comparing these results with state-of-the-art DKT models reported in [12], we observe that iDKT achieves the best predictive performance on the AS2009 and Algebra2005 datasets and ranks second on the Bridge2006 and NIPS34 datasets, surpassed only by the AKT model.

6.4. Research Questions

The experimental validation of iDKT is guided by the following research questions:

1. **Interpretability Validation (RQ1)**: Is it possible to rigorously validate that a iDKT model, whose representations are grounded in a reference model, actually yields interpretable constructs?
2. **Predictive Performance–Interpretability Trade-Off (RQ2)**: To what extent can deep knowledge tracing models be constrained for interpretability alignment without significantly degrading predictive performance?
3. **Higher Granularity (RQ3)**: Can the proposed model capture individualized latent factors at a higher level of granularity than the reference model?

6.5. Interpretability Validation (RQ1)

6.5.1. Alignment Metrics

To verify that iDKT's internal representations—specifically Personalized Initial Knowledge ($\mathbf{l}_c$) and Personalized Learning Rate ($\mathbf{t}_c$) as defined in Equations 3 and 4—faithfully represent the educational constructs postulated by the reference model, we employ two metrics widely utilized in psychometrics and educational measurement [17,18]. To calculate these validation metrics, the latent factors are projected into the probabilistic space $[0,1]$ via the average of the latent dimensions followed by a sigmoid activation:

$$l_u = \sigma(\bar{l}_c), \quad t_u = \sigma(\bar{t}_c) \tag{13}$$

where $\bar{l}_c$ and $\bar{t}_c$ denote the mean across the feature dimension of the individualized embeddings. Verification relies on two metrics:

1. *Convergent Validity (Latent Fidelity)*: This metric, denoted as H_1, is calculated as the Pearson correlation (r) between the projected latent factors (l_u, t_u) and the reference BKT parameters (L_0 and T). High alignment proves the model has successfully internalized the theoretical constructs. The metrics are expressed as:

$$H_1 = \text{Corr}(l_u, L_0) \quad \text{and} \quad H_1 = \text{Corr}(t_u, T) \tag{14}$$

2. *Predictor Equivalence (Behavioral Alignment)*: This metric, denoted as H_2, assesses the functional substitutability of iDKT parameters by executing a *cross-model simulation*. We "inject" the projected iDKT parameters (l_u and t_u) into the canonical BKT recurrence equations to generate *induced mastery trajectories*. This allows us to verify whether the learned factors preserve their causal roles. The metric is statistically quantified as the Pearson correlation coefficient between these induced trajectories and the theoretical reference trajectories generated by the original BKT model:

$$H_2 = \text{Corr}\Big(P(L)_{ind}, P(L)_{ref}\Big) \tag{15}$$

where $P(L)_{ind}$ and $P(L)_{ref}$ represent the sequences of mastery probabilities at each timestep for the induced and reference models, respectively.

Table 2 shows interpretability metrics for the unconstrained iDKT model ($\lambda_{ref} = 0$).

Table 2. Interpretability Alignment Metrics for unconstrained iDKT ($\lambda_{ref} = 0$).

Dataset	H_1 (Init.)	Int.	H_1 (Rate)	Int.	H_2	Int.
AS2009	-0.1382	Poor	-0.0067	Negl.	0.1870	Poor
AS2015	-0.0392	Negl.	0.0908	Negl.	0.0975	Negl.
Algebra2005	-0.0602	Negl.	-0.0273	Negl.	0.0600	Negl.
Bridge2006	-0.0645	Negl.	0.0160	Negl.	0.0809	Negl.
NIPS34	0.2070	Poor	-0.0425	Negl.	0.1722	Poor

Upon enabling Representational Grounding ($\lambda_{ref} = 0.10$), we observe a significant restoration of semantic alignment, as detailed in Table 3.

Table 3. Interpretability Alignment Metrics for grounded iDKT ($\lambda_{ref} = 0.1$).

Dataset	H_1 (Init.)	Int.	H_1 (Rate)	Int.	H_2	Int.
AS2009	0.5166	Good	0.2864	Poor	0.2178	Poor
AS2015	0.9217	Excell.	0.8801	Excell.	0.1749	Poor
Algebra2005	0.5444	Good	0.3310	Fair	0.0661	Negl.
Bridge2006	0.4561	Fair	0.6422	Good	0.0859	Negl.

6.5.2. Discussion of Alignment Results

The results presented in Tables **??** and 3 reveal a notable divergence between *Convergent Validity* (H_1) and *Predictor Equivalence* (H_2). Although H_1 attains high levels across most datasets even with relatively low λ_{ref} grounding weights, H_2 remains consistently low.

High H_1 scores demonstrate that iDKT successfully internalizes the *semantic identity* of the theoretical constructs (Initial Knowledge and Learning Rate). However, H_2 measures functional substitutability within a comparatively constrained reference model that is unable to capture the intricate behavioral patterns identified by iDKT. As a high-capacity Transformer, iDKT captures complex, long-range dependencies and context-aware dynamics that exceed the modeling capability of classical BKT. This discrepancy provides empirical evidence that iDKT does not merely mimic the reference model, but rather maps its core pedagogical constructs onto a more sophisticated and predictive architecture.

6.6. Accuracy–Interpretability Trade-Off (RQ2)

A key contribution of this paper is the systematic exploration of the trade-off between predictive accuracy and interpretability. By modulating the grounding weight λ_{ref} in Equation 7, we identify the Pareto frontier of the model.

Tables 4 and 5 summarize the performance and alignment results for the ASSISTments 2009 and 2015 datasets across a systematic sweep of the grounding weight λ_{ref} (Equation 7). To construct the Pareto frontier, we utilize the *Composite Alignment* $\bar{H}$ as the primary metric for interpretability, defined as the arithmetic mean of the semantic (H_1) and behavioral (H_2) values.

Table 4. Pareto Sweep results for ASSISTments 2009.

λ_{ref}	Test AUC	Composite $\bar{H}$	Interpretation
0.0 (Baseline)	0.8358	0.0140	Black Box
0.1 (Sweet Spot)	0.8290	0.3402	Accurate & Balanced
0.2	0.8096	0.3986	Theory-Dominant
0.3	0.7917	0.4131	Peak Alignment
0.5	0.7686	0.1866	Saturated
1.0	0.7456	0.0378	Theoretical Collapse

Table 5. Pareto Sweep results for ASSISTments 2015.

λ_{ref}	Test AUC	Composite $\bar{H}$	Interpretation
0.0 (Baseline)	0.7252	0.0497	Black Box
0.1 (Sweet Spot)	0.7233	0.6589	99.7% AUC Retained
0.2	0.7169	0.6709	Peak Alignment
0.3	0.7093	0.6639	Saturated
0.5	0.6942	0.6362	Saturated
1.0	0.6724	0.5519	Theory-Locked

The resulting Pareto curves are visualized in Figures 2 and 3. We observe that theoretical guidance serves as a powerful regularizer; even a small weight ($\lambda = 0.10$) causes a massive leap in interpretability with minimal impact on accuracy. However, excessive grounding leads to a "Theory-Locked" state or even "Theoretical Collapse" in denser datasets like AS2009, where the model can no longer reconcile complex behavioral patterns with simple pedagogical assumptions.

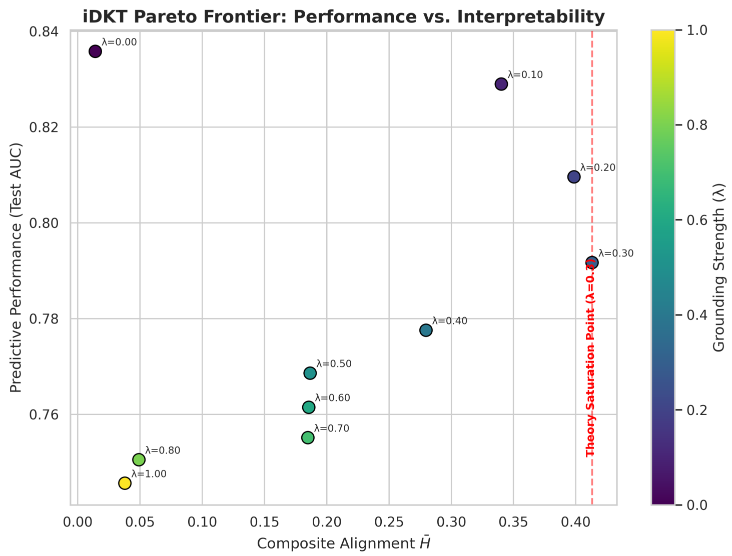


Figure 2. iDKT Pareto Frontier for ASSISTments 2009.

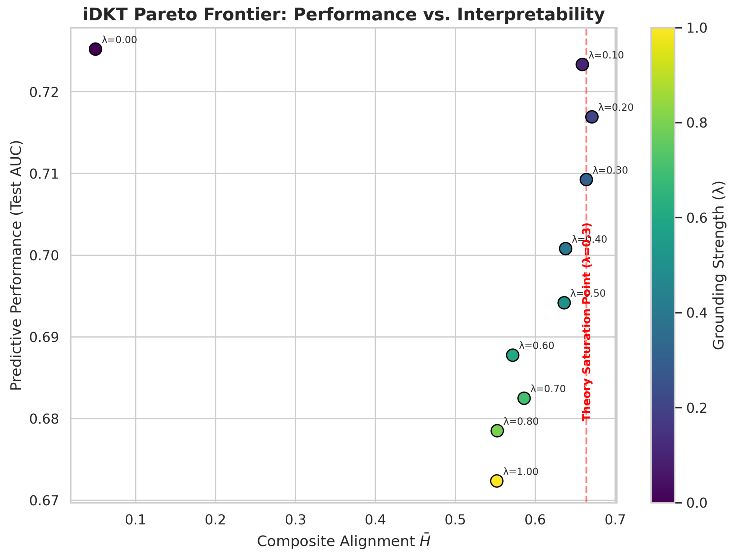


Figure 3. iDKT Pareto Frontier for ASSISTments 2015.

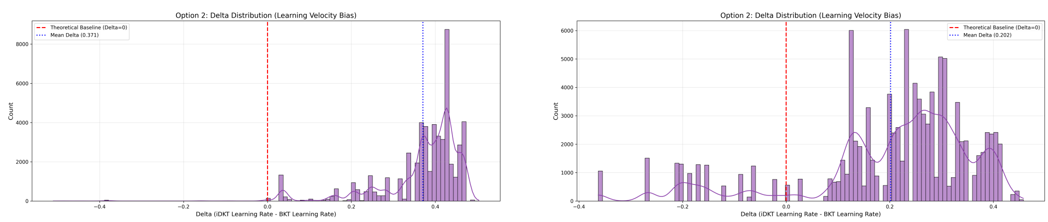
6.7. Higher Granularity (RQ3)

The primary value of iDKT lies in its ability to transform population-level theoretical averages into high-granularity learner diagnostics. While standard BKT assumes that every student shares a fixed "Initial Mastery" (L_0) and "Learning Rate" (T) for a given skill, iDKT decomposes these into individualized traits.

6.7.1. Delta Distribution

To statistically quantify this granularity, we can analyze the "Delta Distribution" ($\Delta = t_s - T$) for the Learning Velocity across all skills. In both ASSISTments sets, we observe a significant non-zero standard deviation ($\sigma \approx 0.019$ for AS2009 and $\sigma \approx 0.039$ for AS2015). This represents a high increase in resolution compared to the BKT baseline, where $\sigma \approx 0$. The right-skewed tail in the distributions identifies a sub-population of "Fast Learners" whose true acquisition pace is systematically underestimated by classical models.

Figure **??** visualizes these densities for both datasets. The centering of the distributions near zero confirms that the model remains effectively anchored to the theoretical priors established by Representational Grounding. However, the substantial width of the curves—the *Individualization Volume*—quantifies the pedagogical information that is lost when using population-level averages. This variance allows the system to distinguish between students who are naturally faster than the prior and those who require more reinforcement to reach mastery.



(a) ASSISTments 2009 delta distribution (b) ASSISTments 2015 delta distribution

Figure 4. Distribution of student learning velocity deviations from the theoretical baseline.

6.7.2. Longitudinal Mastery Dynamics

As shown in Figure 5, iDKT allows estimating distinct trajectories where "Fast" learners reach mastery thresholds significantly earlier than their peers, even when their initial interaction outcomes are similar. This proves that iDKT does not merely mimic BKT but leverages its deep learning capacity to dynamically adjust the velocity of mastery based on the learner's historical profile, enabling truly adaptive pacing.

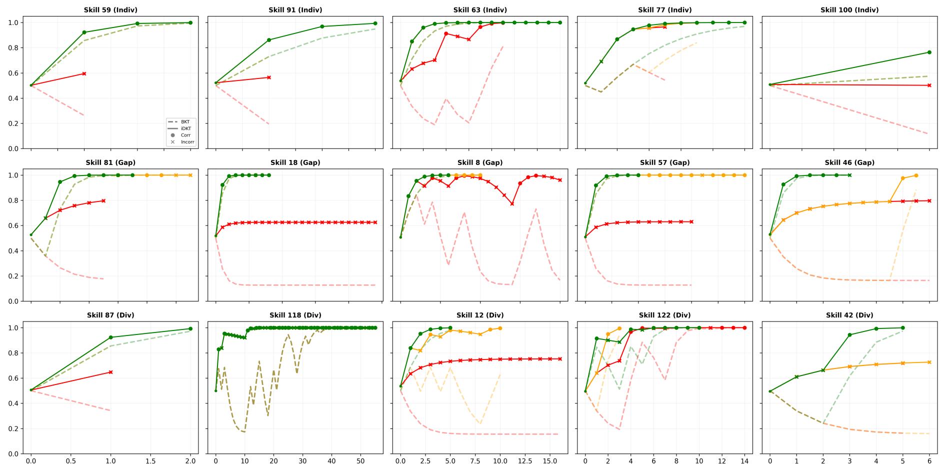


Figure 5. Mastery Alignment Mosaic for ASSISTments 2009. Solid lines represent individualized iDKT trajectories while dashed lines show the population BKT prior.

7. Conclusions

This work has introduced iDKT, a Transformer-based Knowledge Tracing model that bridges the gap between the high predictive capacity of deep learning and the intrinsic interpretability of models such as BKT, which utilize constructs with clear semantic meanings. By utilizing Representational Grounding, we have demonstrated that it is possible to anchor deep latent representations to semantically meaningful constructs without sacrificing state-of-the-art performance.

Our experimental results successfully address the research questions. First, we proved that iDKT successfully internalizes the theoretical constructs of the reference model, achieving high convergent validity (H_1). Second, we formalized a methodology for measuring the trade-off between predictive performance and interpretability, enabling the identification of the "Interpretability Sweet Spot" where significant alignment is achieved with minimal loss in predictive AUC. Finally, we demonstrated that iDKT provides a high increase in diagnostic granularity compared to population-level baselines by identifying individualized learning velocities that enable truly adaptive pacing.

The primary contribution of this research is twofold: a robust methodology for evaluating the internal interpretability of Transformer-based models in education, and a practical architecture that transforms "black-box" predictors into interpretable tools for knowledge tracing. By grounding deep learning in established educational concepts, iDKT offers a path toward AI-driven personalization that is both highly accurate and pedagogically actionable, providing a foundation for next-generation intelligent tutoring systems.

Author Contributions: Conceptualization, C. L.; methodology, C. L. and O. C. S.; software, C. L.; validation, C. L. and O. C. S.; formal analysis, C. L.; investigation, C. L.; resources, C. L. and O. C. S.; data curation, C. L.; writing—original draft preparation, C. L.; writing—review and editing, C. L. and O. C. S.; visualization, C. L.; supervision, O. C. S.; project administration, O. C. S.; funding acquisition, O. C. S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this paper can be downloaded through the links provided in https://pykt-toolkit.readthedocs.io/en/latest/datasets.html.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* **1994**, *4*, 253–278.
2. Šarić Grgić, I.; Grubišić, A.; Gašpar, A. Twenty-five years of Bayesian knowledge tracing: a systematic review, 2022.
3. Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.J.; Sohl-Dickstein, J. Deep knowledge tracing. *Advances in neural information processing systems* **2015**, *28*.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
5. Abdelrahman, G.; Wang, Q.; Nunes, B. Knowledge tracing: A survey. *ACM Computing Surveys* **2023**, *55*, 1–37.
6. Bai, X.; et al. A Survey of Explainable Knowledge Tracing, 2024.
7. Fantozzi, M.; et al. The Explainability of Transformers - Current Status and Directions. *arXiv preprint arXiv:2401.09202* **2024**.
8. Di Marino, S.; et al. Ante-Hoc Methods for Interpretable Deep Models: A Survey, 2025.
9. Karpatne, A.; Atluri, G.; Faghmous, J.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering* **2017**, *29*, 2318–2331.
10. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating Physics-Based Modeling With Machine Learning: A Survey, 2022.
11. Von Rueden, L.; Mayer, S.; Beckh, K.; Georgiev, B.; Giesselbach, S.; Heese, R.; Kirsch, B.; Pfrommer, J.; Pick, A.; Ramamurthy, R.; et al. Informed Machine Learning – A Taxonomy and Survey of Integrating Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering* **2021**, p. 1–1.
12. Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; Tang, J.; Luo, W. pyKT: a python library to benchmark deep learning based knowledge tracing models. *Advances in Neural Information Processing Systems* **2022**, *35*, 18542–18555.
13. Feng, M.; Heffernan, N.; Koedinger, K. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction* **2009**, *19*, 243–266.
14. Stamper, J.; Niculescu-Mizil, A.; Ritter, S.; Gordon, G.; Koedinger, K. Algebra I 2005-2006. *Development data sets from KDD Cup* **2010**.
15. Stamper, J.; Niculescu-Mizil, A.; Ritter, S.; Gordon, G.; Koedinger, K. Bridge to algebra 2006-2007. *Development data sets from KDD Cup* **2010**.
16. Wang, Z.; Lamb, A.; Saveliev, E.; Cameron, P.; Zaykov, Y.; Hernández-Lobato, J.M.; Turner, R.E.; Baraniuk, R.G.; Barton, C.; Jones, S.P.; et al. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061* **2020**.
17. Campbell, D.T.; Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin* **1959**, *56*, 81.
18. Association, A.E.R.; Association, A.P.; on Measurement in Education, N.C.; et al. Standards for educational and psychological testing. *(No Title)* **1985**.