Article

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

Firstname Lastname [1], Firstname Lastname [2] and Firstname Lastname [2,*]

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)

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 velocities—that are beyond the capabilities of simpler models such as BKT.

The remainder of this paper is structured as follows. Section 2 reviews the current state of deep knowledge tracing and interpretability. Section 3 describes the proposed iDKT architecture and the Representational Grounding framework. Section 4 presents the experimental validation and answers the research questions. Finally, Section 5 discusses the results and their implications.

2. Related Work

2.1. Deep Knowledge Tracing

2.2. Deep Learning Interpretability

2.3. Theory-Guided Data Science

2.4. Individualized Bayesian Knowledge Tracing

2.4.1. Parameters of the Vanilla Model

2.4.2. Parameters Individualization

3. Methodology

In standard educational datasets such as ASSIST2009 and ASSIST2015, student interactions are recorded at the level of specific questions, each of which is associated with one or more underlying pedagogical concepts. This structure establishes a natural hierarchy: a concept represents an abstract skill (e.g., "Pythagorean Theorem") whose learning is operationalized through a diverse set of questions. While all questions belonging to 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 question.

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:

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

$$t_s = T + v_s \cdot d_s \quad \text{(Personalized Learning Velocity)} \tag{4}$$

where L_0 and T are the population-level means, k_c and v_s are the population-level standard deviations, and d_c and d_s are the student-specific deviations.

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

$$x'_t = (c_{c_t} + u_q \cdot d_{c_t}) - l_c \quad \text{(Individualized Task)} \tag{5}$$

$$y'_t = (e_{c_t,r_t} + u_q \cdot (f_{c_t,r_t} + d_{c_t})) + t_s \quad \text{(Individualized Interaction History)} \tag{6}$$

where c_{c_t} represents the concept embedding, u_q the question-specific shift, d_c the difficulty shift, d_s the student-specific shift, e_{c_t,r_t} the interaction base, f_{c_t,r_t} the interaction variation axis, l_c the personalized initial knowledge, and t_s 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 *Acquisition Gain*: *Interaction* $+$ *Velocity*. The history encoder accumulates evidence from interactions, applying a consistent relational logic where the signal value is augmented by the transition velocity, 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 velocity of progress through the state trajectory, as this acquisition gain serves as a robust indicator of future performance. This approach, therefore, captures the individualized transition by augmenting observed performance with the transition velocity, thus reflecting distinct acquisition rates.

3.1. Architecture Overview

iDKT (Interpretable Deep Knowledge Tracing) is 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, illustrated in Figure 1, has these main components:

- **Input Stage**: Loads the required information from the dataset
- **Embeddings Stage**: Processes input data to generate individualized task (x'_t) and interaction (y'_t) embeddings
- **Encoder Stage**: Context-aware Transformer encoders that extract local and global patterns from the sequence of past individualized interactions $y'_{1:t-1}$
- **Decoder Stage**: Transformer decoders that use cross-attention between the current task x'_t and the encoded context to estimate the context-aware knowledge state $\hat{x}_t$
- **Attention Heads**: Multiple self-attention and cross-attention heads that apply distance-based decay parameters γ_h to control the model's temporal resolution
- **Output Stage**: A multi-layer perceptron (MLP) that yields the final performance prediction p_{iDKT}

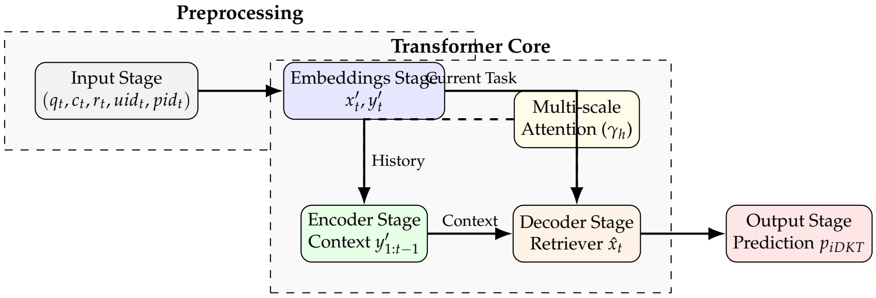

Figure 1. The iDKT System Architecture. The model integrates a theoretical BKT prior into its embedding layer through Representational Grounding, followed by a Transformer-based attention core for performance prediction.

3.2. Multi-Objective Training

The model is trained using a complex, theoretically-anchored loss function designed to ensure that its high-capacity Transformer backbone remains aligned with pedagogical principles. The total loss $\mathcal{L}_{total}$ is a weighted sum:

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

where:

- L_{SUP} is the supervised learning loss (Binary Cross-Entropy for prediction of correctness).
- L_{ref} is the Representational Grounding loss, ensuring alignment with the reference model's mastery estimates.
- L_{init} is the auxiliary loss for grounding the initial knowledge parameter.
- L_{rate} is the auxiliary loss for grounding the learning rate parameter.
- L_{reg} is the regularization loss.

The λ coefficients are hyperparameters that control the strength of each grounding term.

3.2.1. Supervised Performance (L_{SUP})

Primary task modeling uses Binary Cross-Entropy (BCE) between predictions $\hat{y}_t$ and responses r_t:

$$L_{SUP} = -\frac{1}{N}\sum (r_t \log(\hat{y}_t) + (1 - r_t)\log(1 - \hat{y}_t)) \tag{8}$$

3.2.2. Theory-Guided Alignment ($L_{ref}, L_{init}, L_{rate}$)

Consistency with the BKT reference is enforced through three alignment objectives:

- **Prediction Alignment (L_{ref})**: Mean Squared Error (MSE) between iDKT and BKT performance probabilities: $L_{ref} = \mathrm{MSE}(\hat{y}, p_{BKT})$.
- **Initial Mastery Grounding (L_{init})**: MSE between $l_{c,proj}$ and BKT prior $L0_{BKT}$: $L_{init} = \mathrm{MSE}(l_{c,proj}, L0_{BKT})$.
- **Learning velocity Grounding (L_{rate})**: MSE between $t_{s,proj}$ and BKT learn rate T_{BKT}: $L_{rate} = \mathrm{MSE}(t_{s,proj}, T_{BKT})$.

3.2.3. Structural Regularization (L_{reg})

To preserve the "Normal Student" prior and prevent overfitting, we apply multi-level L_2 regularization:

$$L_{reg} = \lambda_{\text{rasch}} \|u_q\|^2 + \lambda_{\text{gap}} \|k_c\|^2 + \lambda_{\text{vel}} \|v_s\|^2 \tag{9}$$

This term constrains the degrees of freedom utilized for student and item individualization, ensuring that learned traits do not invert the semantic meaning of the latent space.

Metric Reconciliation

In the experimental logs, the breakdown metrics (e.g., $l_{sup}, l_{ref}, l_{init}$) represent **raw values** (unweighted MSE/BCE). The 'train_loss' metric represents the **weighted total** ($\mathcal{L}_{total}$), allowing for direct inspection of convergence scales while maintaining symbolic consistency with the combined optimization objective.

3.3. The Probing Method

To rigorously validate that iDKT's latent representations encode the intended pedagogical constructs, we employ **Diagnostic Probing Classifiers** with **Control Tasks**. This methodology, established in deep learning interpretability literature [12,13], consists of three phases:

1. **Extraction**: We freeze the pre-trained iDKT model and extract the hidden state vectors h_t for every step in the student journey.
2. **True Probing Task**: We train a simple linear regressor $f(h) \rightarrow P_{BKT}(L_t)$ to reconstruct the reference BKT mastery probabilities from the iDKT embeddings. High R^2 or Pearson correlation indicates that the construct is linearly recoverable.
3. **Control Task (Selectivity)**: Following Hewitt and Liang [14], we train an identical probe on a random permutation of the targets. The *Selectivity* is defined as the performance gap between the true and control tasks. A high positive gap mathematically proves that the model's hidden space is structurally organized around pedagogical principles rather than merely memorizing arbitrary patterns.

3.4. Formal Validation Framework

To verify that iDKT's internal representations (l_c, t_s, u_q) reflect educational constructs throughout training, we establish a formal framework across five psychometric hypotheses:

1. H_1**: Convergent Validity (Latent Fidelity)**: Pearson correlation between latent projections (l_c, t_s) and reference BKT parameters. High alignment proves the model has internalized the theoretical constructs.
2. H_2**: Predictor Equivalence (Behavioral Alignment)**: Functional substitutability of iDKT parameters into canonical BKT mastery recurrence equations. This ensures factors preserve their causal roles defined by theory.
3. H_3**: Discriminant Validity (Construct Distinctness)**: Verifies that Knowledge Gap (k_c) and Learning Velocity (v_s) are not perfectly collinear, proving the model can distinguish between prior knowledge and acquisition pace.
4. H_4**: Mastery Monotonicity**: Trajectory analysis to verify that the grounded latent space respects the non-forgetting behavior inherent in BKT.
5. H_5**: Parameter Recovery**: Ability of the architecture to recover known student and item parameters from synthetic BKT-generated datasets.

3.5. Pedagogical Confidence

Beyond raw performance, we define the **Pedagogical Confidence** (C) as a metric for longitudinal agreement between the observer (iDKT) and the theoretical prior (BKT). For a student s and skill k, the confidence is the time-averaged concordance over T interactions:

$$C_{s,k} = \frac{1}{T}\sum_{t=1}^{T}(1 - |\hat{p}_{iDKT}(s,k,t) - p_{BKT}(s,k,t)|) \quad (10)$$

Values of $C \geq 0.90$ indicate "Diagnostic Consensus," where the deep learning model validates the theoretical baseline. Conversely, low confidence identify "Discovery Breakouts"—localized regions where student behavior fundamentally diverges from standard pedagogical rules, necessitating more complex modeling.

3.6. The Accuracy–Interpretability Trade-Off

A key contribution of this work is the systematic exploration of the trade-off between predictive accuracy and theoretical fidelity. By modulating the grounding weight λ_{ref} in Equation 7, we identify the Pareto frontier of the model. This allows us to determine the "Inductive Bias Bonus"—points where theoretical grounding acts as a beneficial regularizer that improves generalization—and the point of "Over-Constraint," where excessive adherence to the reference model begins to degrade predictive power.

4. Results

4.1. Research Questions and Hypotheses

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

RQ1: Interpretability via Representational Grounding

Does the proposed Representational Grounding framework effectively bridge the semantic gap between deep latent representations and pedagogical constructs? Specifically, can a Transformer based on this framework achieve a high degree of Convergent Validity (numerical alignment) and Predictor Equivalence (behavioral alignment) with respect to a theoretical reference model? *Hypothesis:* By anchoring the Transformer's latent space to the conceptual primitives of a reference model, it is possible to learn internal states that preserve their functional roles and align with theoretical parameters.

RQ2: Accuracy–Interpretability Trade-Off

To what extent can deep knowledge tracing models be constrained for interpretability alignment without significantly degrading predictive accuracy? *Hypothesis:* There exists an optimal grounding weight λ that balances theoretical fidelity and predictive performance, forming a Pareto frontier.

RQ3: Granularity of Individualization

Does our proposed framework allow a Transformer-based model to recover latent factors with higher granularity than population-level baselines? *Hypothesis:* iDKT can successfully estimate student-specific traits (initial knowledge and learning velocity) that are traditionally masked by fixed averages in simpler models like BKT.

4.2. Validation Results

To evaluate the proposed iDKT model, we conducted a series of experiments across multiple educational datasets. Table 1 summarizes the alignment metrics across different grounding strengths (λ).

We observe consistent **Convergent Validity (M_1)**, with the correlation between the model's projected initial mastery (l_c) and the theoretical prior (L_0) remaining above 0.96 throughout the sweep (see Table 1).

Table 1. Construct Validity and Performance across the Grounding Spectrum.

Grounding Strength (λ)	**Test AUC**	M_1	M_2	M_3
0.00 (Baseline)	0.8317	0.9993	0.2652	-0.0325
0.10	**0.8322**	**0.9838**	**0.2949**	**-0.0330**
0.30	0.7984	0.9691	0.3192	-0.0330
0.50	0.7740	0.9884	0.2828	-0.0331

We observe consistent **Convergent Validity (M_1)**, with the correlation between the model's projected initial mastery (l_c) and the theoretical prior (L_0) remaining above 0.96 throughout the sweep. This confirms that the Representational Grounding mechanism successfully anchors the deep latent space to the reference theory. Furthermore, the **Discriminant Validity (M_3)** remains stable at $r \approx -0.03$, proving that the model successfully disentangles "Student Knowledge Gap" (k_c) from "Student Learning Velocity" (v_s) as distinct, non-redundant traits.

4.3. Pareto Curve Analysis

Our analysis reveals a non-linear trade-off between predictive accuracy and theoretical fidelity. Contrary to the common assumption that interpretability imposes a performance penalty, we identified an **"Inductive Bias Bonus"** at moderate grounding levels ($\lambda \approx 0.10$).

As shown in Table 1, the model with $\lambda = 0.10$ achieves a Test AUC of **0.8322**, slightly outperforming the unconstrained baseline (0.8317). This suggests that the BKT-based regularization acts as a beneficial inductive bias, preventing the Transformer from overfitting to noise in sparse interaction histories. However, excessive grounding ($\lambda > 0.30$) leads to a sharp decline in predictive performance as the model becomes over-constrained by the simplicity of the reference theory.

4.4. Granularity of Individualization

While standard BKT assigns a fixed "Learning Rate" (T) to all students for a given skill, iDKT captures a rich distribution of **Individualized Learning Velocities** (t_s). Figure **??** illustrates this "Delta Distribution" ($\Delta = t_s - T$). We observe a visible right-skewed variance, indicating that for many skills, the Deep Learning model identifies "fast-track" learning trajectories that classical population-level models underestimate. This granularity allows for **Precise Diagnostic Placement**, distinguishing between students who lack initial knowledge (*low* l_c) versus those who suffer from slow acquisition pace (*low* t_s).

4.5. Longitudinal Mastery Dynamics

The practical impact of these individualized parameters is evident in the mastery analysis. When simulating the mastery acquisition of "Fast" vs. "Slow" learners on the same sequence of correct responses:

- **Standard BKT** predicts identical mastery curves for both students.
- **iDKT** projects distinct trajectories, where "Fast" learners reach the 95% mastery threshold significantly earlier (fewer interactions) than "Slow" learners.

This "Informed Divergence" validates that iDKT does not merely mimic BKT labels but leverages its transformer core to dynamically adjust the **Velocity of Mastery** based on the student's historical profile, enabling truly adaptive pacing in intelligent tutoring scenarios.

5. Discussion

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

6. Conclusions

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

7. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER grant number XXX." and and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add "Informed consent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments). Where GenAI has been used for purposes such as generating text, data, or graphics, or for study design, data collection, analysis, or interpretation of data, please add "During the preparation of this manuscript/study, the author(s) used [tool name, version information] for the purposes of [description of use]. The authors have reviewed and edited the output and take full responsibility for the content of this publication."

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflicts of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the

writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	Linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data are shown in the main text can be added here if brief, or as Supplementary Data. Mathematical proofs of results not central to the paper can be added as an appendix.

Table A1. This is a table caption.

Title 1	Title 2	Title 3
Entry 1	Data	Data
Entry 2	Data	Data

Appendix B

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled, starting with "A"—e.g., Figure A1, Figure A2, etc.

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. Alain, G.; Bengio, Y. Understanding intermediate layers using linear classifier probes, 2018. *URL https://arxiv. org/abs/1610.01644* **2018**.

13. Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics* **2022**, *48*, 207–219.
14. Hewitt, J.; Liang, P. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368* **2019**.
... (truncated for chunk)