MDPI

Article

Beyond Black-Box Deep Knowledge Tracing: Structurally Grounded Transformers 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

While deep knowledge tracing models provide high predictive accuracy, their black-box nature limits the extraction of actionable pedagogical insights. This study introduces iDKT, an interpretable-by-design Transformer model that utilizes *structural grounding* to align deep latent representations with educational constructs defined by intrinsically interpretable models. 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 it 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 population-level baselines provided by the reference model. Specifically, iDKT identifies student-level initial knowledge and learning velocities, providing mastery estimations that are more sensitive to the nuances of individual behavioral patterns than those produced by standard BKT. By anchoring deep learning to semantic concepts defined by the reference model, iDKT enables precise diagnostic placement and dynamic pacing in adaptive learning environments. 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 mining; personalized learning

1. Introduction

Knowledge Tracing (KT) is a fundamental task in the fields of Artificial Intelligence in Education, Intelligent Tutoring Systems (ITS) and Massive Open Online Courses (MOOCs). 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 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 [1,2], relies on probabilistic graphical models that explicitly represent latent knowledge states. BKT is intrinsically interpretable: its parameters (e.g., probability of learning, initial knowledge, slipping, guessing) map directly to pedagogical constructs, allowing educators to diagnose student difficulties

Received:
Revised:
Accepted:
Published:

Copyright: © 2025 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.

and trust the model's decisions. However, its simplicity often limits its predictive power, struggling to capture the complex, non-linear dependencies inherent in real-world learning trajectories [**?**].

The second paradigm emerged with the advent of Deep Knowledge Tracing (DKT) [**?**], which utilizes Recurrent Neural Networks (RNNs) and later Transformers [**?**] to model student interactions as complex sequential data. Models such as DKT, DKVMN [**?**], and AKT [**?**] have achieved state-of-the-art predictive performance, significantly outperforming classical approaches by capturing latent long-term dependencies. 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 educational theory. 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 they derive actionable pedagogical insights from the model's internal weights [3].

Current efforts to bridge this gap typically rely on post-hoc explainability methods, such as visualization of attention weights or perturbation analysis [4,5]. 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 valid educational constructs.

To address this challenge, we propose a shift towards Interpretability-by-Design, inspired by the emerging paradigm of Theory-Guided Data Science (TGDS) [6,7]. TGDS posits that scientific consistency should be an architectural constraint rather than an afterthought. By integrating extensive domain knowledge—in this case, pedagogical theory—deep learning models can be constrained to learn representations that are both scientifically plausible and highly predictive.

In this work, we introduce iDKT (Interpretable Deep Knowledge Tracing), a novel Transformer-based model that achieves intrinsic interpretability through Structural Grounding. Unlike previous approaches that use theory only for regularization [8], iDKT is designed to anchor its latent representations directly to the conceptual space of a reference model. This allows to leverage the power of Transformers to capture complex learning dynamics while ensuring that its internal states remain formally equivalent to established educational parameters.

Our contributions are threefold: (1) We propose a method for Structural Grounding that forces a Transformer to maintain internal states anchored in semantically meaningful constructs; (2) We validate iDKT against BKT, demonstrating that it captures granular, student-specific insights—such as individualized knowledge gaps and diverse learning velocities—that classical models overlook; and (3) We show that this approach enables actionable educational interventions, such as Precise Diagnostic Placement and Dynamic Pacing, without sacrificing the predictive accuracy characteristic of state-of-the-art deep learning.

2. Methodology

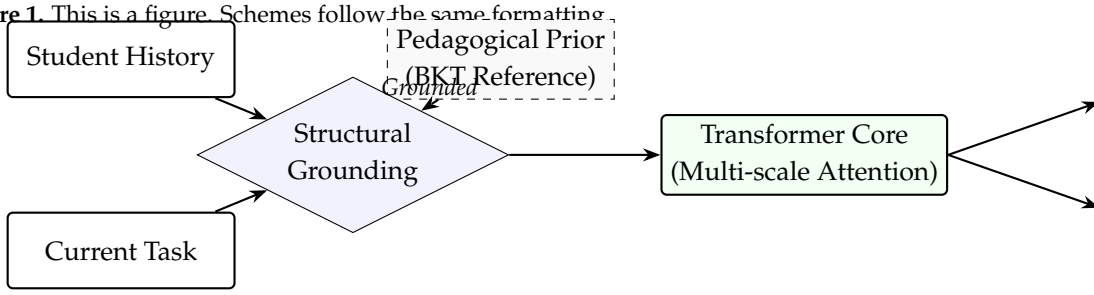
2.1. The iDKT model

iDKT (Interpretable Deep Knowledge Tracing) is a Transformer-based model designed to bridge the gap between predictive power and pedagogical interpretability. It builds

upon the attention mechanisms of an encoder-decoder architecture [**?**] introducing a novel input layer based on structural grounding. The core architecture consists of three main components:

- Context-Aware Encoder: Processes the sequence of exercises and questions to generate contextualized question embeddings.
- Decoder: A structure that retrieves relevant historical interactions using an attention mechanism. The multi-head attention employs distinct, learnable decay rates for each head to capture short-term and long-term dynamics, ensuring a comprehensive view of the student's learning trajectory.
- Prediction Head: A multi-layer perceptron (MLP) that combines the multi-head output with the current question embedding to predict the probability of a correct response.

Figure 1. This is a figure. Schemes follow the same formatting



2.2. Structural Grounding Embeddings

To achieve interpretability-by-design, iDKT replaces standard learned embeddings with "Textured Grounding" embeddings. These are formally anchored to the conceptual space of Bayesian Knowledge Tracing (BKT) [1], ensuring that the latent representations carry semantic meaning. We employ a modified Rasch model logic to individualize these representations:

- **Individualized Question (x'_t)**: The standard question embedding is replaced by a residual representation:
$$x'_t = (c_{c_t} + u_q \cdot d_{c_t}) - l_c$$
where c_{c_t} is the concept embedding, u_q is a learned scalar for item difficulty, and d_{c_t} is a variation axis. Crucially, l_c is the **individualized initial mastery**, grounded in the BKT prior ($L0$), defined as:
$$l_c = L0_{skill} + k_c \cdot d_c$$
Here, k_c is a learned student-specific scalar representing their "Knowledge Gap" relative to the population mean.
- **Individualized Interaction (y'_t)**: The interaction history is similarly grounded by adding learning momentum:
$$y'_t = (e_{c_t,r_t} + u_q \cdot (f_{c_t,r_t} + d_{c_t})) + t_s$$

where t_s is the **individualized learn rate**, grounded in the BKT learn probability (T):

$$t_s = T_{skill} + v_s \cdot d_s$$

Here, v_s represents the student's "Learning Velocity," allowing the model to distinguish between fast and slow learners dynamically.

2.3. Loss Functions and Training Objective

The model is trained using a multi-objective loss function that balances predictive accuracy with theoretical alignment. The total loss L_{total} is a weighted sum of three components:

$$L_{total} = L_{SUP} + \lambda_{ref} L_{ref} + L_{reg}$$

- **Supervised Prediction Loss (L_{SUP})**: The standard Binary Cross-Entropy (BCE) loss between the predicted probability $\hat{p}_{it}$ and the actual student response r_{it}.
- **Theoretical Alignment Loss (L_{ref})**: Enforces consistency with the reference theory (BKT). It includes Mean Squared Error (MSE) terms penalizing deviations between the model's projected parameters (e.g., l_c, t_s) and the corresponding BKT theoretical values. This ensures that the learned representations remain semantically valid.

$$L_{ref} = \text{MSE}(l_c, L0_{BKT}) + \text{MSE}(t_s, T_{BKT})$$

- **Regularization Loss (L_{reg})**: A task-agnostic regularization on the student-specific scalars (u_q, k_c, v_s) to prevent overfitting and ensure they represent meaningful deviations from the norm.

3. Results

3.1. Validation Metrics

To verify that iDKT's internal representations genuinely reflect educational constructs rather than arbitrary latent features, we evaluated the model against three validation metrics (H_1–H_3).

3.1.1. M1: Convergent Validity (Numerical Alignment)

- **Statement**: The model's latent projections (l_c, t_s) exhibit a high Pearson correlation ($r > 0.90$) with the intrinsic parameters of the reference BKT model.
- **Support**: Drawing on the **Informed Machine Learning** paradigm [9] and the **TGEL-Transformer** framework [10], which emphasize numerical alignment between neural components and theoretical rules.
- **Demonstration**: Alignment metrics (`initmastery_corr`, `learning_rate_corr`) calculated on the test set.

3.1.2. M2: Predictor Equivalence (Behavioral Alignment)

- **Statement**: The iDKT parameters (l_c, t_s) are **functionally substitutable**; when plugged into the reference BKT equations, they reconstruct a mastery trajectory that is highly consistent with the reference model's behavior.
- **Support**: This follows the **Structural Grounding** principle: for a parameter to represent a construct, it must not only correlate with it (H1) but also fulfill its causal/functional role in the reference theory's equations.
- **Methodology**: Calculate $\hat{y}_{induced,t} = \text{BKT}(l_{c,idkt}, t_{s,idkt}, s_{bkt}, g_{bkt})$ and measure its correlation with BKT baseline outputs.

- **Demonstration**: Functional alignment correlation > 0.60.

3.1.3. M3: Discriminant Validity (Construct Distinctness)

- **Statement**: The student-specific knowledge gap (k_c) and learning velocity (v_s) capture **non-redundant** dimensions of variance, proving they represent distinct pedagogical features even if they exhibit natural positive correlation.
- **Support**: In psychometrics, discriminant validity does not imply zero correlation (empirical independence), but rather that the two constructs are not **perfectly collinear**. If $r(k_c, v_s) \approx 1.0$, the model would suffer from an **identifiability problem**, where it couldn't distinguish if a correct response is due to "knowing more" or "learning faster."
- **Demonstration**: Correlation analysis showing $r(k_c, v_s) < 0.85$, ensuring that each parameter provides a unique contribution to the marginalized accuracy. This allows for the identification of "high-velocity/low-prior" students (under-prepared but fast learners) vs. "low-velocity/high-prior" students (well-prepared but struggling with new acquisition).

Table 1 summarizes the alignment metrics across different grounding strengths (λ).

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 Structural 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.

3.2. 3.2. Pareto Curve

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.

3.2.1. 3.3. 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 1 (see supplementary materials) 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 momentum (*low* t_s).

3.2.2. 3.4. Longitudinal Mastery Dynamics

The practical impact of these individualized parameters is evident in the **Mastery Mosaic** 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.

4. Results 2

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

4.1. Subsection

4.1.1. Subsubsection

Bulleted lists look like this:

- First bullet;
- Second bullet;
- Third bullet.

Numbered lists can be added as follows:

1. First item;
2. Second item;
3. Third item.

The text continues here.

4.2. Figures, Tables and Schemes

All figures and tables should be cited in the main text as Figure **??**, Table 1, etc.

Figure 2. This is a figure. Schemes follow the same formatting.

Table 2. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Title 1	Title 2	Title 3
Entry 1	Data	Data
Entry 2	Data	Data [1]

The text continues here (Figure 3 and Table 3).

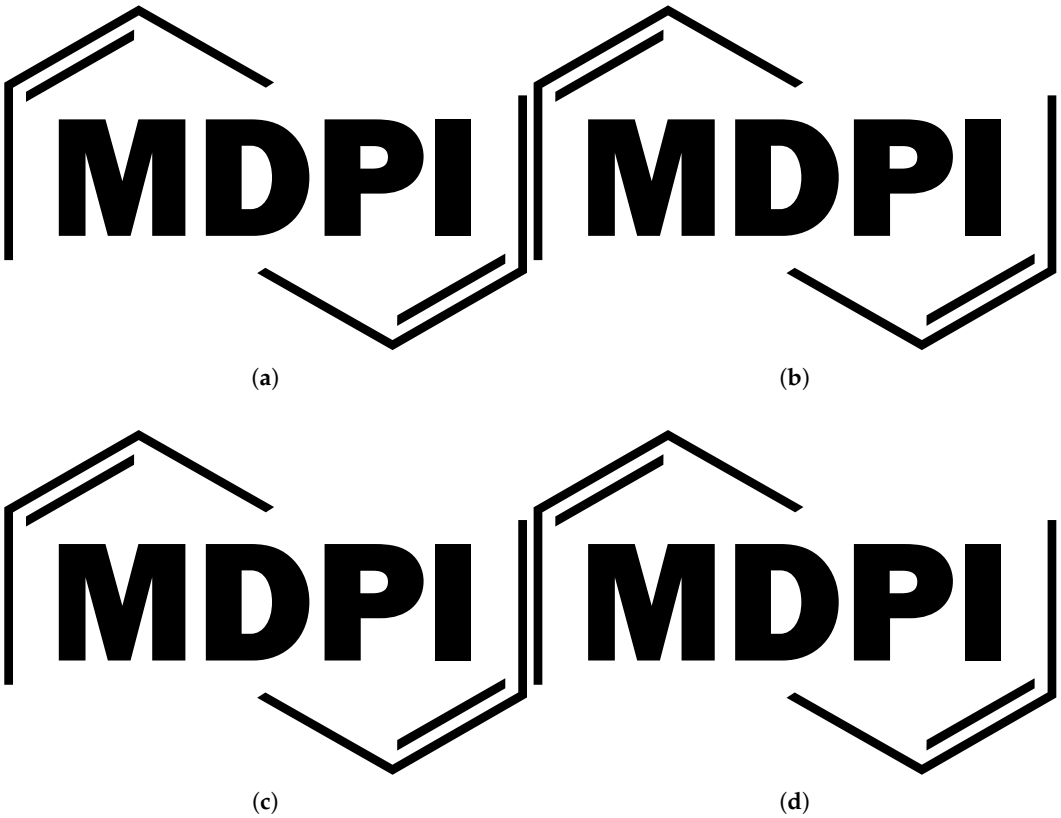


Figure 3. This is a wide figure. Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. (**c**) Description of what is contained in the third panel. (**d**) Description of what is contained in the fourth panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

Table 3. This is a wide table.

Title 1	Title 2	Title 3	Title 4
Entry 1 *	Data Data Data	Data Data Data	Data Data Data
Entry 2	Data Data Data	Data Data Data	Data Data Data

* Tables may have a footer.

Text.

Text.

4.3. Formatting of Mathematical Components

This is the example 1 of equation:

$$a = 1, \tag{1}$$

the text following an equation need not be a new paragraph. Please punctuate equations as regular text.

This is the example 2 of equation:

$$a = b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z \quad (2)$$

Please punctuate equations as regular text. Theorem-type environments (including propositions, lemmas, corollaries etc.) can be formatted as follows:

Theorem 1. *Example text of a theorem.*

The text continues here. Proofs must be formatted as follows:

Proof of Theorem 1. Text of the proof. Note that the phrase "of Theorem 1" is optional if it is clear which theorem is being referred to. □

The text continues here.

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. Yudelson, M.V.; Koedinger, K.R.; Gordon, G.J. Individualized bayesian knowledge tracing models. In Proceedings of the International conference on artificial intelligence in education. Springer, 2013, pp. 171–180.
3. Bai, X.; et al. A Survey of Explainable Knowledge Tracing, 2024.
4. Fantozzi, M.; et al. The Explainability of Transformers - Current Status and Directions. *arXiv preprint arXiv:2401.09202* **2024**.
5. Di Marino, S.; et al. Ante-Hoc Methods for Interpretable Deep Models: A Survey, 2025.
6. 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.
7. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating Physics-Based Modeling With Machine Learning: A Survey, 2022.
8. Lee, U.; Park, Y.; Kim, Y.; Choi, S.; Kim, H. Consistency and monotonicity regularization for neural knowledge tracing. *arXiv preprint arXiv:2106.06965* **2021**.
9. 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.
10. Gong, Y.; et al. TGEL-transformer: Fusing educational theories with deep learning for interpretable student performance prediction. *Expert Systems with Applications* **2025**.

... (truncated for chunk)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.