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:

comes at the cost of a simplicity that often limits its predictive power, making them struggle to capture the complex, non-linear dependencies existent 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 they derive actionable pedagogical insights from the model's internal weights [6].

Current efforts to bridge this gap typically rely on post-hoc explainability methods, such as visualization of attention weights 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.

1.1. Research Questions

The specific research questions that guided our experimental validation are as follows.

Research Question 1: Interpretability Through 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?

We hypothesize that by anchoring the Transformer's latent space to the conceptual primitives of a reference model, it is possible to learn internal states that not only are numerically correlated with theoretical values but also preserve their functional roles. We evaluate this through two primary probing metrics: (i) the correlation between projected latent factors and parameters of the reference model, and (ii) the ability of these projected factors to reconstruct behavioral trajectories when substituted into the reference model's equations, thereby confirming that the latent representations are semantically anchored and that the internal decision-making process adheres to the underlying logic of the reference model.

Research Question 2: Accuracy vs. Interpretability Trade-Off

To what extent can deep knowledge tracing models be constrained for interpretability alignment without significantly degrading predictive accuracy?

We hypothesize that it's possible to measure interpretability by means of Representational Grounding and that, by modulating the strength of such grounding, we can quantitatively measure the trade-off between both objectives. We evaluate this by performing a systematic sweep of the grounding weight λ, which allows us to construct a Pareto frontier and identify the optimal trade-off point where predictive accuracy and theoretical fidelity are balanced.

Research Question 3: High-Granularity

Does our proposed *Representational Grounding* framework allow a Transformer-based model to recover values for latent factors that align with the semantic constructs of a theoretically-grounded reference model, providing higher-granularity information?

We hypothesize that iDKT can successfully estimate the latent variance of student-specific traits that are traditionally masked by the fixed population-level averages of simpler approaches. To evaluate this, we take BKT as the reference model, analyze the distribution of individualized initial knowledge and learning rates, and estimate longitudinal mastery dynamics, illustrating how iDKT discovers novel behavioral patterns that remain hidden within the BKT baseline.

The remainder of this paper is structured as follows.

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

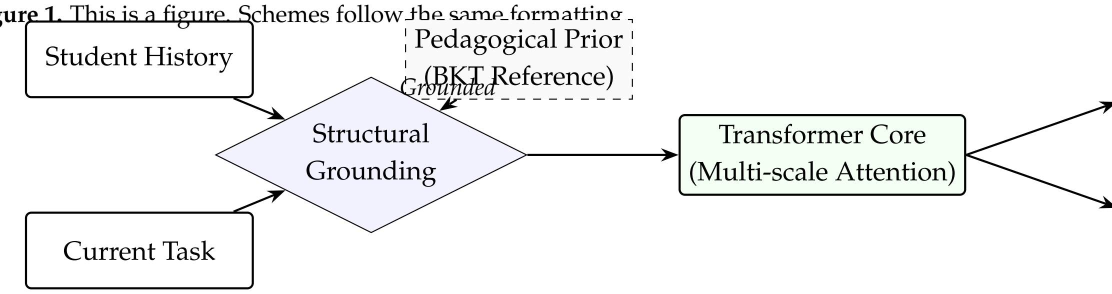
3. Methodology

3.1. 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 [4] introducing a novel input layer based on Representational 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.



3.2. The 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.

3.3. The Probing Method

3.4. 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.

3.5. Interpretability Measure

3.6. The Tradeoff Interpretability vs Prediction Accuracy

4. Results

4.1. 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).

4.1.1. 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 [11] and the **TGEL-Transformer** framework [12], which emphasize numerical alignment between neural components and theoretical rules.
- **Demonstration**: Alignment metrics (`initmastery_corr`, `learning_rate_corr`) calculated on the test set.

4.1.2. 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 **Representational 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.

4.1.3. 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 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.

4.2. 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.

4.2.1. 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).

4.2.2. 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.

5. 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.

5.1. Subsection

5.1.1. 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.

5.2. 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).

(**a**) (**b**)

(**c**) (**d**)

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.

5.3. 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:

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

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.

6. 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.

7. Conclusions

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

8. 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. 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)