Review

Not peer-reviewed version

A Systematic Review of Deep Knowledge Tracing (2015-2025): Toward Responsible AI for Education

Ekaterina Krivich * , Danial Hooshyar , Gustav Šír , Yeongwook Yang , Merja Bauters , Raija Hämäläinen , Tommi Kärkkäinen

Posted Date: 23 October 2025

doi: 10.20944/preprints202510.1845.v1

Keywords: deep knowledge tracing; neural networks; responsible AI for education; systematic review

Review

A Systematic Review of Deep Knowledge Tracing (2015-2025): Toward Responsible AI for Education

Ekaterina Krivich [1,*], Danial Hooshyar [1,2,3], Gustav Šír [4], Yeongwook Yang [5], Merja Bauters [1,6], Raija Hämäläinen [2] and Tommi Kärkkäinen [3]

1 School of Digital Technologies, Tallinn University, Tallinn, Estonia
2 Faculty of Education and Psychology, University of Jyväskylä, Jyväskylä, Finland
3 Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland
4 Department of Computer Science, Czech Technical University, Prague, Czech Republic
5 Department of Computer Science and Engineering, Gangneung-Wonju National University Wonju, Republic of Korea
6 Department of Philosophy, History and Art Studies, University of Helsinki, Helsinki, Finland
* Correspondence: krivich@tlu.ee

Abstract

Background and Objectives: Tracking and adapting to learners' evolving knowledge is essential for effective teaching. In digital learning, Deep Knowledge Tracing (DKT) employs deep neural networks to analyze sequential learner interactions, model their evolving knowledge, and predict skill mastery over time. While DKT is widely studied, their real-world adoption remains limited. This review examines DKT research from 2015–2025 through the lens of responsible AI principles, investigating modeling trends, evaluation practices, input features used for representing learner performance and context, strategies for mitigating data quality issues, assessment of sequential stability (consistency of knowledge estimates over time), and interpretability for educators. **Methods:** Following PRISMA guidelines, five major scholarly databases (Web of Science, Scopus, ScienceDirect, ACM Digital Library, IEEE Xplore) and Google Scholar were searched, yielding 1,047 peer-reviewed articles. Following two rounds of screening and a quality appraisal focused on methodological rigor, 84 studies were included in the final synthesis. **Results:** Graph-based architectures were most common (26.2%), followed by Hybrid/Meta (23.8%) and Attentive models (17.9%). ASSIST datasets were used in 82.1% of studies, and 90.5% predominantly used Area Under the Curve (AUC) for evaluation. A wide variety of input features were used, ranging from basic question–answer pairs and knowledge concepts to time-based metrics, difficulty levels, behavioral indicators, and learning resource interactions. Approaches to address data quality challenges appeared in 44.0% of studies. Only 3.6% quantitatively assessed sequential stability of predictions. Interpretability techniques—designed to make predictions understandable to educators—were present in 11.9% of studies. **Conclusions:** Current DKT models often overlook responsible AI principles, including robust handling of data quality issues, assessment of sequential stability of predictions, and interpretability of predictions. As AI regulatory frameworks increasingly mandate trustworthy and interpretable AI in education, future research should prioritize these principles for practical and responsible deployment.

Keywords: deep knowledge tracing; neural networks; responsible AI for education; systematic review

---

1. Introduction

The growing integration of Artificial Intelligence (AI) in education offers significant opportunities to enhance teaching practices, enrich student learning experiences, and streamline educational management. AI shows potential at both the classroom and system levels—for instance, by personalizing learning for diverse students, identifying those at risk, and assessing emerging competencies [1–5]. Central to most AI systems is the learner model—a computational representation of a student's cognitive and non-cognitive characteristics. Learner models are often built from digital interaction

 doi:10.20944/preprints202510.1845.v1

data (e.g., interactions with learning systems) and provide real-time assessments of learner progress, allowing educators and AI tools to adjust instructions accordingly [6,7].

Two primary AI approaches are used to construct learner models: symbolic and sub-symbolic. Symbolic methods—such as rule-based systems and Bayesian networks—offer interpretable models and support the integration of expert knowledge, but they face challenges with noisy data and the high cost of encoding real-world educational problems into symbolic representations, a limitation known as the knowledge acquisition bottleneck [8]. In contrast, sub-symbolic methods—particularly deep learning models—can automatically learn complex, non-linear patterns from (non)sequential data, reducing reliance on manual feature engineering and enabling more scalable, data-driven learner modeling [9,10].

Deep Knowledge Tracing (DKT) is a prominent sub-symbolic approach that applies recurrent neural networks to capture the temporal dynamics of student knowledge [11]. Since its introduction, DKT has been shown to outperform traditional models in predictive accuracy, yet its advantages come with challenges—particularly opacity, potential bias, and a lack of pedagogical grounding [12–16]. In education, these shortcomings risk reinforcing inequalities and limiting meaningful adaptation to individual learner needs [3,17].

Addressing such issues requires moving beyond purely technical optimization toward responsible AI—a human-centered approach grounded in fairness, transparency, accountability, and educational validity [14,18,19]. As education is increasingly recognized as a high-risk domain for AI deployment (European Union's AI Act [20]), trustworthiness and interpretability become not just technical goals but ethical imperatives for learner modeling. The next sections examine the evolution from classical to deep knowledge tracing, key deployment challenges from a responsible AI perspective, existing reviews, and the need for this study—outlining both the advances achieved and the challenges that remain—followed by the research questions.

1.1. From Classical to Deep Knowledge Tracing

To understand knowledge tracing in the context of learner modeling, it is essential to first clarify what phenomenon and process we are modeling. At the core of this approach lies the concept of knowledge components (KCs)—the fundamental cognitive units representing specific skills, concepts, or competencies that students must acquire and master through practice [21]. These KCs can range from basic arithmetic operations such as addition or subtraction to more complex problem-solving strategies in domains like physics, programming, or language learning. In educational environments, students engage with learning items such as exercises, questions, or problems, each of which requires the application of one or more KCs. As illustrated in Figure 1, the knowledge tracing process involves observing sequences of student interactions with learning items, where each interaction yields an observable response—typically coded as correct or incorrect. However, the actual knowledge state—the degree to which a student has mastered each KC—remains latent and must be inferred from these observable interaction patterns. The central computational challenge in KT is to model how this hidden knowledge state evolves over time as students practice, learn, forget, or consolidate their understanding, and subsequently to predict future performance on new or unseen items based on historical interaction sequences [21,22]. This modeling task is inherently sequential because learning is a temporal process where prior experiences influence current knowledge states, which in turn determine the probability of success on subsequent learning opportunities.

Knowledge Tracing (KT)—one of the most extensively studied approaches for learner modeling—estimates how a learner's mastery of latent knowledge components evolves over a sequence of practice opportunities. Classical KT methods have been foundational in educational data mining since the 1990s. Bayesian Knowledge Tracing (BKT), introduced by Corbett and Anderson [23], models student learning as a Hidden Markov Model with binary latent states (mastered/not mastered) and four parameters: initial knowledge, learning rate, guess, and slip probabilities. Its strengths include strong theoretical grounding in cognitive science, interpretable parameters aligned with educational theories, and computational efficiency. However, the original BKT formulation suffers from several limitations: the binary mastery assumption oversimplifies continuous learning processes [24,25], the standard model does not incorporate forgetting nor handle multiple skills simultaneously [23,26], and parameter identifiability issues can lead to degenerate solutions where different parameter combinations yield identical predictions [24,27,28]. While numerous extensions have been developed to address these limitations, including forgetting-aware variants and multi-skill formulations, these enhancements often increase model complexity and computational requirements (e.g., [2,29]).

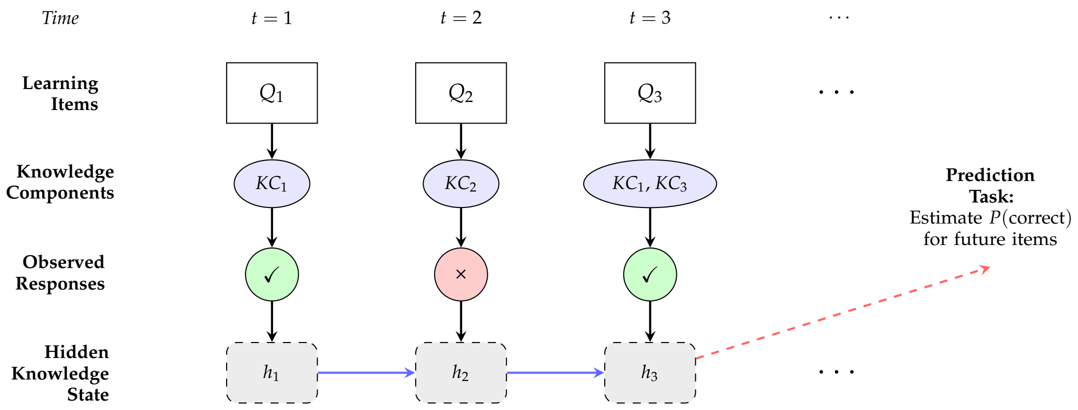

Figure 1. The knowledge tracing process. Students interact with learning items over time, each associated with one or more knowledge components (KCs). Observed responses (correct ✓ or incorrect ×) are used to infer the hidden knowledge state, which evolves sequentially. The goal is to predict future performance based on this latent state trajectory. Adapted from Abdelrahman et al. [21].

Performance Factor Analysis (PFA) by Pavlik et al. [29] addresses some BKT limitations by using logistic regression to predict performance based on prior successes and failures, allowing for more granular skill modeling. While PFA offers greater flexibility and can handle multiple skills, it still relies on hand-crafted features, assumes linear relationships between practice and performance, and cannot capture complex temporal patterns or student-specific learning trajectories [9,30]. Other classical approaches like Additive Factor Model (AFM) and Item Response Theory (IRT) variants face similar trade-offs between interpretability and expressiveness, often requiring extensive domain expertise for feature engineering and struggling to capture the non-linear, heterogeneous nature of real student learning [31–33]. These limitations stem from the need for sophisticated parameter tuning and feature selection processes that demand substantial educational domain knowledge [34], while their linear assumptions fail to adequately model the complex cognitive processes and heterogeneous learning patterns observed in real educational contexts [35,36].

While early work (e.g. BKT, PFA) relied on hand-crafted features and symbolic approaches [9,37], the introduction of DKT by Piech et al. [11] marked a major change by applying recurrent neural networks (specifically LSTMs) to capture complex temporal dependencies in student learning. The term "deep" refers to the use of multi-layer neural networks, and the original DKT specifically employed deep recurrent architectures (LSTMs). Since knowledge tracing is fundamentally a temporal prediction task, numerous other deep learning approaches beyond recurrent architectures have been developed for time series forecasting, including transformers, graph neural networks, and convolutional architectures [38–40]. Notably, transformers avoid recurrence entirely, relying solely on attention mechanisms [41], and have been successfully adapted to knowledge tracing (as discussed later in the Attentive model category). This sub-symbolic approach promised to overcome the limitations of traditional methods through automatic feature learning and improved predictive accuracy [30,42]. Since its introduction, numerous studies have focused on improving DKT's prediction accuracy and addressing its limitations [22,43]. However, despite its ability to predict, DKT has revealed significant methodological challenges. Yeung and Yeung [44] identified critical issues with consistency over time: the model sometimes predicts lower mastery after correct answers and produces sequentially inconsistent predictions that weaken the assumption of gradual knowledge acquisition. Xiong et al. [13]

showed that DKT does not perform much better than simpler models like PFA on prepared datasets, while also revealing weaknesses in handling multi-skill sequences and raising concerns about dataset quality. Wang et al. [45] further analyzed DKT's instability through finite-state automata, showing that prediction volatility stems from built-in structural limitations rather than merely noisy data.

Additionally, DKT faces challenges common to deep neural networks: limited ability to integrate prior domain knowledge, susceptibility to learning spurious correlations from biased data (e.g., imbalanced datasets), and lack of interpretability [14,46]. Subsequent variants, such as self-attentive or graph-based KT, have improved predictive accuracy [47,48]. Yet, these technical refinements have not fully resolved underlying issues related to stability and transparency. The persistence of these problems—despite a decade of research—highlights the need for a critical examination of DKT research.

1.2. Critical Problems in DKT Deployment from Responsible AI Perspective

The concept of responsible AI has been articulated through various frameworks, each emphasizing key principles such as fairness, privacy, accountability, and transparency. For instance, Maree et al. [49] highlight fairness, privacy, accountability, transparency, and soundness, while Arrieta et al. [50] expand this view to include ethics, security, and safety. Other perspectives, such as those of Eitel-Porter [51] and Werder et al. [52], stress explainability as a crucial component. A broader perspective is offered by Jakesch et al. [53], incorporating sustainability, inclusiveness, social good, human autonomy, and solidarity into the responsible AI discourse. Building on these foundations, Goellner et al. [54] define responsible AI as a human-centered approach that builds user trust through ethical decision-making, explainable outcomes, and privacy-preserving implementation.

Synthesizing these contributions, this review examines DKT through the lens of responsible AI, drawing on prior critical reviews and surveys that link DKT research to key principles found across responsible AI frameworks. These include dealing with data quality issues, ensuring sequential stability so that pedagogical decisions remain reliable over time, and transparency—operationalized through process-level interpretability to support educator trust and regulatory compliance. Poor data quality (e.g., data noise, imbalanced distributions, sparsity) can undermine fairness by producing biased or inequitable learner assessments that disadvantage certain student groups or learning trajectories [55–57]. Overlooking sequential stability undermines accountability by leading to inconsistent or unreliable knowledge predictions, potentially resulting in inappropriate instructional actions [14,15,58]. A lack of interpretability risks educator distrust and non-compliance with oversight standards, whereas clear, process-level explanations strengthen transparency and help justify AI-driven decisions [59]. Together, these dimensions operationalize fairness, accountability, and transparency in the deployment of DKT, forming the foundation for responsible and trustworthy use in educational settings.

1.2.1. Data Inconsistency and Class Imbalance

Educational logs are inherently imbalanced: some items are trivially easy while others are rarely answered correctly [55,60]. This imbalance can cause DKT models to overfit to dominant patterns in the data, weighting their predictions toward frequently occurring or easier items. As a result, the models may misrepresent proficiency for learners who primarily encounter less-represented or more challenging items, indirectly creating biased assessments across different groups of students. Evidence of this emerges when mainstream knowledge tracing models are re-evaluated on resampled, balanced test sets: they show significant performance drops, revealing a heavy reliance on answer distribution biases rather than genuine learning patterns [56]. This bias is particularly pronounced in minority skills—knowledge components with limited representation in training data or extreme success/failure rates. Models may incorrectly classify students as proficient in rarely practiced skills based solely on overall performance patterns, or fail to detect genuine learning progress in challenging skills, leading to inappropriate instructional decisions and inequitable educational outcomes.

1.2.2. Sequential Stability

Sequential stability refers to a model's ability to produce consistent, educationally plausible trajectories of a learner's knowledge state over time. This construct encompasses several related aspects: *temporal consistency* (maintaining coherent predictions across time steps), *smoothness* (avoiding abrupt oscillations in mastery estimates), and *educational plausibility* (producing trajectories that align with learning theory). The lack of sequential stability creates serious problems for practical tutoring systems, as educators require mastery estimates that change smoothly and plausibly over time to make informed instructional decisions [44,45]. Current DKT models exhibit "wavy transitions" and dramatic fluctuations in knowledge state predictions even when students provide consistent responses, sometimes predicting lower mastery after correct answers [14,61–63]. These unstable trajectories prevent educators from trusting the system's assessments, undermine actionable feedback, and may indicate problematic over-fitting that renders models unreliable for real-world deployment [44].

1.2.3. Process-Level Interpretability

The black-box nature of deep learning models creates significant barriers to adoption in educational contexts where educators demand explanations for AI-driven assessments [59,64]. This lack of interpretability prevents educators from understanding why the system makes specific predictions, undermining trust and potentially leading to harmful automated decisions. The absence of interpretable models also creates regulatory compliance problems, as the European Union's AI Act requires explainability for high-risk AI systems in education.

While explainable AI methods such as LIME and SHAP exist [5,65], many current DKT models overlook implementing any interpretability mechanisms, creating a fundamental mismatch between technological capability and practical educational needs. It is important to distinguish between interpretability and explainability, terms that are often used interchangeably in the literature but represent distinct approaches to model transparency [16,66]. Interpretability refers to the degree to which a human can understand the cause of a decision through the inherent design of the model—achieved through architectural choices that make the decision-making process inherently transparent [67]. Explainability, in contrast, often refers to the ability to provide post-hoc explanations for model decisions through external methods applied after training, such as attention visualization or feature attribution techniques [68–70]. This distinction becomes crucial in educational contexts where both the immediate comprehensibility of model behavior (interpretability) and the ability to justify specific decisions to stakeholders (explainability) are essential for practical deployment and regulatory compliance. However, recognizing that many DKT studies in the literature use these terms interchangeably, this review adopts a pragmatic approach by classifying studies based on the actual techniques employed rather than the terminology used. Because the conceptual frameworks scholars cite can ultimately influence the techniques they select, resolving this misalignment remains a critical issue; establishing a clear view of what is currently implemented provides the necessary groundwork.

Collectively, these problems create urgent challenges for educational practice that require systematic evaluation. Educators need reliable, fair, and interpretable models to make informed instructional decisions, yet current DKT research fails to ensure these fundamental requirements are met. This necessitates an organized synthesis that looks beyond typical features and characteristics of DKT studies and examines how they address existing challenges to achieve responsible AI principles.

1.3. Existing Reviews and the Need for This Study

Despite the growing interest in (deep)KT, existing reviews have predominantly concentrated on model taxonomies and performance comparisons, offering limited insight into the deeper methodological challenges that affect the real-world deployment of these models. For instance, Dai et al. [6] conducted a review of KT techniques from the perspectives of assumptions, data, and algorithms, highlighting that most models address only a subset of assumptions about knowledge components and cognitive processes, and predominantly rely on quiz data as input. They identified dynamic

 doi:10.20944/preprints202510.1845.v1

Bayesian networks, logistic regression, and deep learning as the main algorithmic approaches. Song et al. [22] discussed different aspects of KT models to identify their distinctions and better support research in the field. Their review provided a granular categorization of mainstream models, detailed analyses of techniques and technological solutions, and outlined potential future research directions in deep learning–based KT. Shen et al. [71] conducted a comprehensive survey of KT, categorizing three fundamental model types, reviewing their variants under stricter learning assumptions, and showcasing typical application scenarios. They also introduced two open-source libraries—EduData for dataset access and preprocessing, and EduKTM for unified model implementation—to support research and practice, and discussed future development directions in the field. Bai et al. [72] provided a comprehensive review of explainable knowledge tracing (xKT), introducing core concepts from both xAI and KT, and categorizing KT models into transparent and black-box types. They reviewed interpretability methods across ante-hoc, post-hoc, and other dimensions, highlighted the lack of robust evaluation methods, and demonstrated three xAI approaches on ASSISTment2009, offering insights for improving interpretability evaluation from an educational stakeholder perspective. Abdelrahman et al. [21] provide a survey of KT models, covering methods from early approaches to recent deep learning–based techniques, while highlighting theoretical aspects, benchmark dataset characteristics, key modeling differences, and current research gaps, as well as outlining possible future research and application directions.

While these syntheses have undoubtedly contributed to advancing the (deep) knowledge tracing field by organizing research, guiding model development, and documenting methodological trends, notable gaps remain. No existing work has provided a systematic review of DKT that comprehensively collects and analyzes all relevant studies from its inception to the present. Crucially, most reviews focus on methodological taxonomies and performance comparisons but overlook practical challenges—such as data quality issues, sequential instability, and lack of interpretability—that are critical for ensuring the responsible AI compliance and trustworthiness of DKT systems. The current landscape therefore lacks an evidence-based synthesis that interrogates not only which models perform well, but also how they operate under real-world constraints that matter to educators and learners. This study addresses these gaps by presenting the first systematic review of DKT to explicitly evaluate methodological rigor in alignment with responsible AI principles, providing a foundation for advancing both technical performance and ethical readiness in educational AI.

1.4. Research Questions

This study synthesizes the DKT literature from 2015–2025 to identify critical issues that undermine reliable deployment and practical utility in educational settings. To address these problems and guide responsible AI–aligned deployment, the research questions are:

RQ1 (Modeling landscape): Which datasets, model architectures, input features (learner state indicators and context features), and validation metrics have been employed in DKT research, and how do these features distribute across the current taxonomy?

RQ2 (Data inconsistency): How do studies prevent biased student assessments by recognizing and mitigating data inconsistency issues in their predictive modeling?

RQ3 (Sequential stability): How do studies ensure reliable pedagogical decisions by evaluating the sequential stability of their models in relation to the evolution of knowledge?

RQ4 (Process-level interpretability): How do studies enable educator trust and regulatory compliance by providing interpretability for their DKT models?

2. Method

To design and implement this systematic review, we followed the guidelines outlined by PRISMA's framework [73].

2.1. Database and Keywords

We conducted a thorough search across five major scholarly databases to ensure that our review was as comprehensive and unbiased as possible. These databases include Web of Science, Scopus, ScienceDirect, ACM Digital Library, and IEEE Xplore. Additionally, we explored grey literature by examining the first ten pages of Google Scholar [74] and selected the most pertinent research articles based on our predefined criteria. Our method involved employing keyword combinations identified through existing related works. The core search terms were formulated as ("knowledge tracing") AND (deep* OR neural*) and used to search within the title, abstract, and keywords, adapted to each database's syntax requirements.

2.2. Eligibility Criteria

We established specific criteria to filter relevant articles. These criteria were divided into inclusion and exclusion categories, with inclusion criteria set in advance to guide our initial search. These included criteria such as publications being in English, peer-reviewed, and dated from 2015 to 2025. Exclusion criteria, refined during screening, removed studies without empirical evaluation, limited to classical knowledge tracing methods, or lacking sufficient methodological detail. For a detailed breakdown of these criteria, see Table 1.

Table 1. Criteria for paper eligibility.

Inclusion criteria
I1 The year of publication includes studies from 2015 to 2025
I2 Publications from conferences or peer-reviewed journals
I3 Written in English and full text accessible
I4 Focuses on deep knowledge tracing models
Exclusion criteria
E1 Study without evaluations: lacks validation or empirical studies
E2 Conference publications outside the main conference (e.g., short papers, posters), theses, or inaccessible full text
E3 The study focuses exclusively on classical (non-deep) knowledge tracing methods
E4 Applications of existing DKT frameworks without novel contributions
E5 Studies with insufficient methodological detail regarding model architecture or evaluation
E6 Studies not adequately addressing methodological challenges in educational contexts

2.3. Study Selection Process

Figure 2 illustrates the comprehensive PRISMA flow diagram illustrating the sequential stages of study identification, screening, and final selection. The article selection process comprised three primary steps of identification, screening, and eligibility evaluation. After the search, we imported our search results into Zotero reference management software and conducted both automatic and manual duplicate searches to eliminate duplications. Records identified from databases (n = 958) and Google Scholar (n = 89) were combined, with 311 duplicates removed.

In the first screening phase, the primary researcher evaluated abstracts and titles against the eligibility criteria, assigning scores of 0 (exclude), 0.5 (unsure), or 1 (advance to full-text screening). Another researcher independently coded a randomly selected subset to verify consistency with the primary researcher's assessments. They also independently re-evaluated all abstracts initially scored as 0 and provided independent ratings for those marked as 0.5. Uncertain cases progressed to the second screening phase, while the first screening excluded 542 records with discrepancies resolved through consensus discussion.

In the second screening phase, 194 full texts were assessed following the same methodological approach, with the primary researcher conducting initial evaluations while the another researcher independently reviewed a subset and verified all articles scored 0 or 0.5. Additional exclusion criteria were refined during this stage as needed. Of the 194 full-text articles assessed, 5 could not be retrieved due to access limitations (E2), and 105 papers were excluded for specific reasons: lacking validation or empirical studies (E1, n = 29), classical KT only (E3, n = 32), no novel contribution (E4, n = 20; referring to comparative or applied studies using existing DKT frameworks without introducing methodological or conceptual innovations), insufficient methodological detail (E5, n = 17), and poster papers (E2, n = 7). All discrepancies were resolved through discussion until consensus was reached, resulting in 84 studies included in the final synthesis.

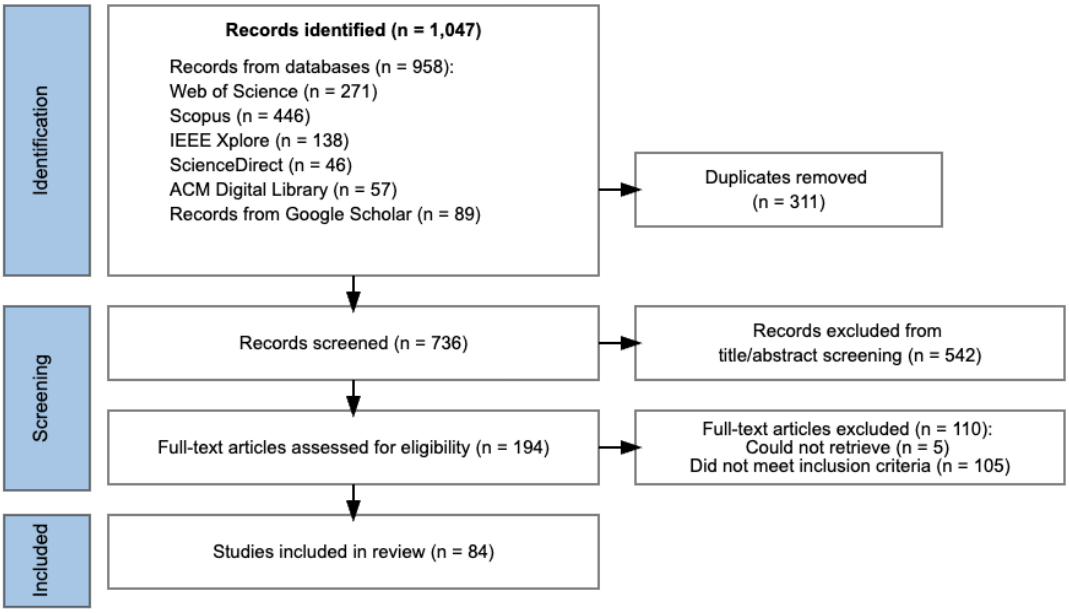

Figure 2. Study selection procedure.

2.4. Quality Appraisal

Quality assessment is a pivotal stage in systematic literature reviews to ensure reliability and mitigate bias in study findings. Following the methodologies from previous reviews and guidelines [75], this review employs a tailored set of five criteria that emphasize the significance of internal and external validity, as well as the relevance and appropriateness of the research methodology and its execution. The five quality appraisal criteria are:

(1) Are the objectives and the context of research clear, and well connected with DKT?
(2) Does the research adequately delineate the methodology (e.g., training dataset, the model architecture, and its hyper-parameters)?
(3) Does the research adequately delineate the experiments (e.g., examining sequential or temporal consistency, assessing interpretability, addressing data inconsistencies, or evaluating the model's real-world applicability)?
(4) Is there congruence between methodology and methods used for data collection, analysis, and interpretation?
(5) Are the study contributions and limitations clearly stated?

The quality assessment framework allocated points according to predefined criteria: a response of "Yes" indicated high quality (1 point), "To some extent" indicated medium quality (0.5 points), and "No" indicated low quality (0 points). The highest possible score was 5. Studies scoring 3.5 points or higher were classified as high quality, those between 1.5 and 3.5 points as moderate quality, and those scoring 1.5 points or less as low quality. This approach ensured that only studies of sufficient quality were included, thereby supporting valid and reliable conclusions. All included articles received scores above the low-quality threshold, with all 84 papers falling into the medium- or high-quality categories.

2.5. Information Extraction and Data Analysis

Extracted information covered publication details (author names, country, year, publication type, keywords), study characteristics (research objectives, theoretical frameworks, sample), technical specifications (study design, data sources, DKT model architecture, evaluation metrics), methodological rigor indicators (data quality handling, sequential stability assessment, interpretability techniques), and study outcomes (findings, challenges, and future directions). Two researchers independently extracted and categorized this information using a standardized data extraction form.

For the first research question, we applied a deductive content analysis approach informed by the taxonomy of Abdelrahman et al. [21], which we adapted to suit the scope and objectives of this review. First, we reconceptualized "Text-Aware" category as a cross-cutting input modality rather than a standalone architectural category, mapping models that leverage textual features on their dominant computational mechanism (e.g., attention-driven text models are classified as Attentive, graph-enhanced text models as Graph-Based). Second, we introduced the Hybrid/Meta category to accommodate models that integrate multiple architectural paradigms or employ meta-learning approaches, which were not explicitly categorized in the original taxonomy. To ensure clarity and mutual exclusivity, we assigned each model to exactly one primary category based on its most distinctive architectural characteristics. These categorizations were derived from architectural references in the methods sections or, when unspecified, inferred from detailed model descriptions and methodologies in the reviewed articles. In addition to architectures, we systematically coded datasets, input features (including learner state indicators and context features), and validation metrics, drawing on detailed descriptions of model designs and methodologies extracted from the reviewed articles.

To address the second to fourth research questions, we employed an inductive thematic analysis approach to identify patterns in how studies addressed methodological challenges. This process involved three main activities: (1) identifying explicit mentions of data quality issues and the mitigation strategies reported in each study (e.g., missing data, data noise, imbalanced distributions, sparsity, inconsistent timestamps, and other data inconsistencies), (2) cataloguing the methods used to evaluate sequential stability—temporal consistency and smoothness—of model predictions (e.g., visual inspection through heatmaps, prediction trajectory plots, and moving-average smoothing curves), and (3) documenting the interpretability techniques applied (e.g., feature importance analysis, attention weight visualization, rule extraction, counterfactual examples, and concept-based explanations). Emergent themes were then iteratively refined through constant comparison across studies, with any coding disagreements resolved through discussion until consensus was reached. Accordingly, we created tables and visualizations to synthesize findings.

3. Results

Table A1 in the Appendix provides a comprehensive summary of all 84 included studies with their key characteristics. As shown in Figure 3, the included articles originated from eight countries (based on first author's affiliation), with **China** contributing the majority of publications (e.g., IDs 1, 4, 6, 12), followed by **Japan** (e.g., IDs 17, 20, 40), the **USA** (e.g., IDs 14, 63, 84), **Australia** (e.g., ID 16), **Estonia** (e.g., IDs 37, 80), **Hong Kong** (e.g., IDs 44, 58), the **Republic of Korea** (e.g., IDs 23, 71), and **Canada** (e.g., ID 49). Although the initial pool covered more countries, the application of eligibility criteria narrowed the final sample to 84 articles from these eight, with representation across Asia, North America, and Europe highlighting DKT's multi-regional recognition. Figure 4 presents the temporal evolution of DKT studies across model categories from 2015 to 2025, highlighting rapid growth after the model's introduction in 2015, with the highest peak in 2024 (37 papers) driven mainly by **hybrid/meta** (e.g., IDs 22, 31, 79) and **graph-based** models (e.g., IDs 26, 29, 33). The data reveals distinct developmental phases: initial **sequence-based** explorations (2015; e.g., ID 84), diversification into **attention** and **graph** architectures (2018–2020; e.g., IDs 14, 40), and recent convergence toward **hybrid approaches** that combine multiple paradigms (2021–2025; e.g., IDs 22, 31, 79). Finally, Figure 5 illustrates the bibliometric analysis of the included articles, with 33 **conference papers** (e.g., IDs 8, 10, 14) and 51

 doi:10.20944/preprints202510.1845.v1

journal publications (e.g., IDs 1, 3, 17), respectively. The analysis reveals a significant increase in journal publications commencing in 2022 (e.g., IDs 2, 49, 69), a trend that aligns with the accelerated adoption of educational technologies during the global pandemic. **Conference proceedings** have maintained consistent publication rates throughout the review period (e.g., IDs 8, 10, 14), indicating sustained interest and ongoing methodological discourse within the research community.

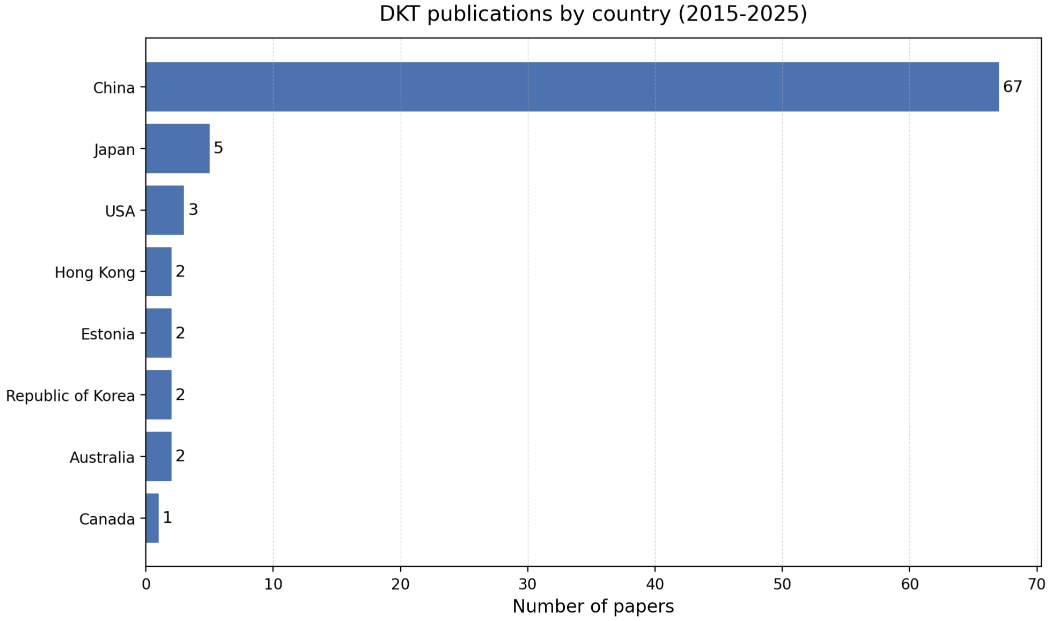

Figure 3. Geographic distribution of DKT research publications by country.

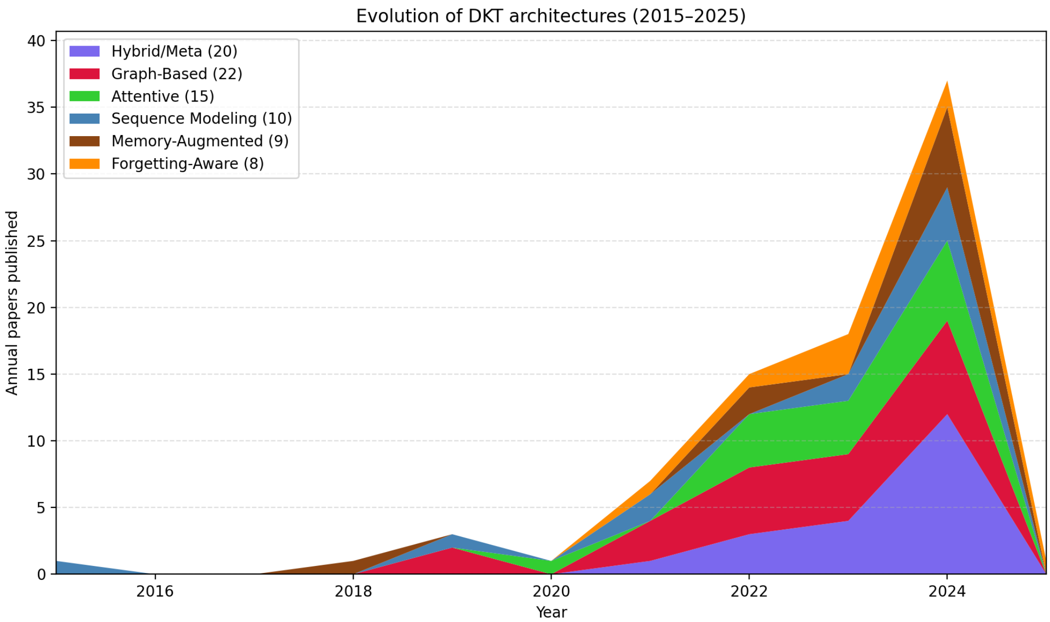

Figure 4. Evolution of DKT model categories. The stacked area chart shows the actual annual distribution of papers across different architectural categories.

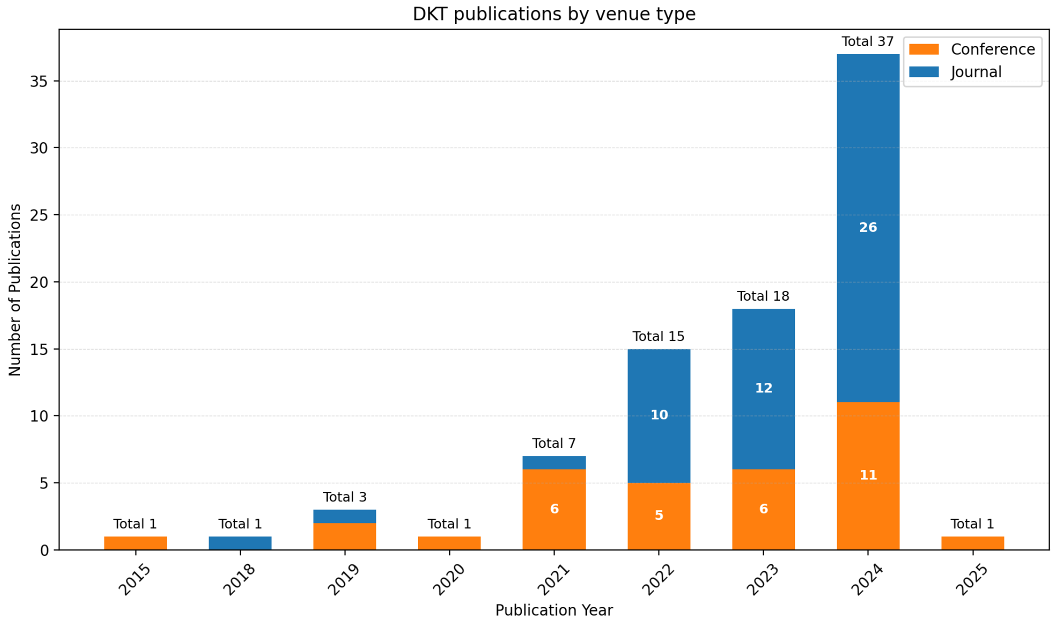

Figure 5. Distribution of publications by type (conference vs. journal) from 2015 to 2025.

3.1. RQ1: Modeling Landscape

The investigation of the current DKT modeling landscape used a four-part analytical framework examining: (1) taxonomic classification of model architectures and their evolution over time, (2) patterns of dataset utilization across the research collection, (3) characteristics of learners and their learning processes, modeled through behavioral indicators and contextual features, and (4) the evaluation metrics employed for model assessment. Application of this taxonomic classification revealed distinct architectural paradigms with varying prevalence in the literature.

Hybrid/Meta models represent a large segment of the reviewed studies, accounting for 23.8% of the literature (e.g., IDs 10, 22), combining multiple paradigms or exploring novel architectures without a single dominant characteristic. **Graph-based** models comprise 26.2% of the studies (e.g., IDs 4, 6), explicitly capturing domain structure through graph neural networks, enabling the modeling of prerequisite relationships and skill dependencies that sequential approaches cannot address, in line with cognitive theories emphasizing the interconnectedness of knowledge [76]. **Attentive** models account for 17.9% of the studies (e.g., IDs 2, 5), employing self-attention mechanisms as the primary temporal modeling component, with categorization requiring that attention weights—not recurrent states—perform most of the temporal modeling. **Sequence Modeling** models represent 11.9% of the studies (e.g., IDs 8, 84) and rely on recurrent architectures without sophisticated attention or external memory, reflecting the foundational DKT paradigm through vanilla RNN/LSTM/GRU cores. **Forgetting-aware** models account for 9.5% (e.g., IDs 16, 77), explicitly incorporating temporal decay mechanisms to reflect cognitive theories of knowledge degradation, requiring mechanisms that reduce past knowledge contributions with elapsed time. **Memory-augmented** models represent 10.7% of the literature (e.g., IDs 25, 44), maintaining explicit external memory structures—key-value or slot-based—for learnable skill representations, which qualitatively differ from implicit hidden state propagation.

Figure 4 illustrates the temporal evolution of these categories from 2015 to 2025, showing the recent surge in DKT research and the current diverse ecosystem. The data reveals the field's rapid growth, with hybrid approaches demonstrating increasing sophistication and graph-based approaches emerging prominently around 2019 in recent years, reflecting the field's maturation and multi-paradigm value recognition.

DKT Data sources. Dataset utilization patterns across the reviewed corpus reveal dominance of ASSIST datasets, employed in 82.1% of the reviewed studies (69 of 84) (e.g., IDs 8, 12, 23), establishing a de facto

 doi:10.20944/preprints202510.1845.v1

standard benchmark while simultaneously introducing potential limitations in generalizability and cross-cultural applicability. This concentration reflects both pragmatic considerations—standardized benchmarks facilitate model comparison—and path dependencies that may constrain innovation in educational data representation.

The ASSIST family of datasets (ASSISTments 2009, 2012, 2015, and 2017) widespread adoption stems from several factors: public availability without restrictive licensing, pre-processed formats requiring minimal data engineering, established baseline performances enabling direct comparison, and comprehensive documentation of problem-skill mappings (e.g., IDs 8, 17). However, this ubiquity creates methodological monoculture where models optimized for ASSIST's specific characteristics—binary correctness labels, single-skill assumptions, and particular student demographics—may fail to generalize to diverse educational contexts. EdNet, utilized in 26.2% of studies (e.g., IDs 5, 62, 71), introduces complexities absent in ASSIST: multi-modal question types, hierarchical content structures, and detailed timestamp granularity enabling fine-grained temporal analysis. Studies employing EdNet demonstrate increased attention to computational efficiency and scalability challenges, though cultural specificity in content and learning patterns limits global applicability. Statics2011, appearing in 14.3% of studies (e.g., IDs 23, 60), emphasizes multi-step problem solving and partial credit scoring that challenge binary classification assumptions prevalent in simpler benchmarks. Junyi, appearing in 8.3% of studies (e.g., IDs 20, 65), contributes large-scale K–12 activity logs commonly used to evaluate generalization beyond ASSIST. Programming judge datasets appear in 2.4% of studies (e.g., IDs 19, 56), highlighting domain-specific challenges such as compilation feedback, iterative debugging, and non-binary outcomes that differ from mathematics tutoring contexts.

The remaining studies employ diverse data sources including game-based learning environments (1.2%; e.g., ID 37) and synthetic datasets (2.4%; e.g., IDs 77, 84). Game-based datasets introduce engagement metrics and voluntary participation dynamics absent in classroom settings, while synthetic data enables controlled experimentation with known ground truth but risks oversimplifying real-world complexity. Notably absent from the corpus are datasets from developing educational contexts, vocational training, or adult learning environments, revealing geographic and demographic blind spots in current DKT research.

This dataset concentration creates cascading effects throughout the research pipeline. Studies exclusively using ASSIST demonstrate lower rates of addressing data quality issues (38%) compared to those employing diverse datasets (47%), suggesting that benchmark familiarity may breed complacency regarding data limitations. Furthermore, the predominance of mathematics and STEM content in available datasets potentially biases architectural innovations toward quantitative reasoning patterns while neglecting humanistic disciplines requiring different cognitive models.

Modeled characteristics. Beyond dataset selection, the reviewed studies exhibit substantial variation in how they represent and model learner characteristics, as detailed in Table 2. Using the inductive analysis and labeling of the methodology section for each paper where authors described features used as inputs to their DKT models, we identified 1–3-word summaries whose prevalence is as follows: **Task performance** (the sequence of correct/incorrect answers) in 52 studies; **Knowledge components/skills** in 40; **Extended interaction logs** (raw learner interaction traces beyond basic sequences) in 17; **Item difficulty** in 16; **Time-related features** (response time or inter-event intervals) in 11; **Textual features** in 7; **Behavioral signals** (e.g., speed, attempts, hints, option choices, or media usage) in 4; and **Knowledge mapping** (mappings of items to skills or multiple skill requirements) in 2. The analysis reveals that most models rely on performance data, often augmented with a subset of temporal, semantic, or difficulty features.

 doi:10.20944/preprints202510.1845.v1

Table 2. Input features used in DKT models.

Theme	Number of Articles	Article IDs
Task performance	52	1, 2, 5, 6, 7, 8, 10, 12, 13, 14, 15, 17, 20, 23, 24, 25, 26, 27, 30, 31, 32, 33, 34, 35, 38, 39, 40, 41, 44, 45, 46, 47, 49, 50, 53, 54, 59, 61, 62, 63, 65, 68, 69, 70, 71, 74, 75, 79, 81, 82, 83, 84
Knowledge components/skills	40	2, 3, 4, 9, 11, 13, 15, 16, 18, 21, 27, 28, 29, 30, 31, 35, 38, 39, 41, 43, 46, 47, 48, 49, 51, 52, 53, 56, 57, 58, 59, 62, 64, 66, 67, 72, 73, 77, 78, 82
Time-related features	11	2, 9, 11, 13, 16, 22, 43, 46, 64, 76, 80
Textual features	7	14, 18, 19, 54, 58, 60, 69
Item difficulty	16	7, 9, 10, 19, 27, 29, 34, 35, 43, 47, 48, 58, 60, 62, 66, 78
Behavioral signals	4	30, 61, 63, 73
Knowledge mapping	2	3, 4
Extended interaction logs	17	1, 3, 4, 7, 11, 22, 28, 36, 38, 40, 51, 52, 55, 67, 77, 78, 80

Evaluation metrics. Our analysis reveals broad agreement on evaluation metrics, with **AUC** being the most common measure, used in 76 studies (90.5%; all paper IDs except: 21, 42, 44, 48, 49, 55, 63, 72), making it the dominant standardfor assessing DKT model performance across diverse educational contexts. **Accuracy** was reported in 48 studies (57.14%; e.g., IDs 3, 10, 20, 22, 23) and **Root Mean Square Error (RMSE)** in 13 studies (15.48%; e.g., IDs 10, 29, 62, 72) as additional metrics. The heavy reliance on AUC reflects the field's focus on binary classification and its consideration of imbalanced classes, though it does not fully capture the temporal dynamics and sequential dependencies that are central to student learning processes (see Section 3.3).

Table 3. DKT model categories with their core deep learning building blocks, representative models, and discriminative cues.

Category	Building-block sub-architecture	Representative KT models	Discriminative cue	Paper IDs
Sequence Modeling	RNN / Long Short-Term Memory (LSTM) / Gated Recurrent Unit (GRU) layers	DKT [11], DKT+ variants	Hidden-state recurrence; no external memory or self-attention.	8, 20, 23, 28, 36, 52, 53, 60, 70, 84
Memory-Augmented	Key–value memory network	DKVMN, BCK-VMN	External differentiable key–value slots updated each step.	25, 35, 44, 47, 48, 67, 68, 73, 78
Attentive	Pure self-attention / Transformer	SAKT, SAINT(+), AKT	Multi-head attention re-weights past items; no recurrence.	2, 3, 5, 7, 9, 11, 12, 14, 15, 18, 38, 56, 61, 74, 81
Graph-Based	Static concept-graph GCN	GKT, JKT	Learner state propagated on a fixed concept graph.	4, 6, 13, 26, 29, 33, 40, 41, 42, 43, 45, 46, 49, 50, 54, 55, 58, 59, 65, 66, 69, 72
Forgetting-Aware	Time-decay gating	LPKT, DKT-Forget, DGMN	Explicit decay functions/gates reduce distant influences.	16, 17, 24, 30, 51, 63, 76, 77
Hybrid/Meta	Ensemble or novel architectures	Mixing-Framework, SAKT+DKT, Graph+Decay	Combines multiple paradigms or explores novel architectures.	1, 10, 19, 21, 22, 27, 31, 32, 34, 37, 39, 57, 62, 64, 71, 75, 79, 80, 82, 83

3.2. RQ2: Data Inconsistency and Bias

The organized examination of data quality issues (e.g., missing and noisy data, imbalanced distributions, sparsity, inconsistent timestamps, and other inconsistencies) across the 84 reviewed manuscripts reveals substantial shortcomings in how these fundamental concerns are methodologically addressed in contemporary DKT research.

Figure 6 presents a parallel sets diagram visualizing the relationships between data quality issues and their corresponding treatment techniques across the reviewed corpus. The majority of studies (56.0%) neglect data quality concerns despite their critical impact on model performance and generalizability. Across the corpus, sparse data is addressed in 22.6% (19/84) of studies, followed by imbalance (15.5% = 13/84) and Q-matrix bias (3.6% = 3/84). Additionally, isolated studies address causal debias (1.2% = 1/84) and data noise (1.2% = 1/84). The figure highlights the diversity of treatment techniques, ranging from embedding/cold-start methods and contrastive learning to preprocessing and graph/relational methods. The width of the connecting flows indicates how issue types map to multiple strategies, showing that researchers often adopt complementary approaches to manage complex data quality challenges. To further unpack these patterns, the following sections examine each category of data quality issues in detail, outlining the specific challenges they pose, and the strategies proposed to address them.

Sparse data (22.6% = 19/84 studies; e.g., IDs 11, 59, 71): This challenge arises when students exhibit limited interaction histories with specific skills or items, constraining reliable estimation of latent knowledge states. Reported mitigation strategies fall into four recurrent design patterns. First, cold-start initialization modules introduce learnable student and item (or skill) embeddings to enable prediction with minimal prior history (e.g., IDs 28, 73). Second, contrastive or comparative representation learning increases robustness under sparsity by pulling semantically or conceptually similar interaction pairs closer while pushing dissimilar ones apart, improving generalization from few observations (e.g., IDs 53, 73). Third, structural or relational propagation methods leverage auxiliary graphs—skill–skill, question–concept, or curriculum relations—to infer unobserved proficiency signals through neighborhood aggregation or counterfactual augmentation (e.g., IDs 41, 65, 69). Fourth, side-information integration enriches sparse sequences with additional modalities such as question text, difficulty or knowledge-structure references, and engineered statistics to densify representations (e.g., IDs 27, 59). Additional preprocessing tactics (e.g., interpolation, padding, sequence truncation, removal of ultra-short histories; IDs 37, 76) and sampling/negative sampling schemes (ID 4) complement these approaches by regularizing extremely short or irregular interaction traces. Collectively, these techniques operationalize sparsity handling through a blend of architectural inductive bias, representation learning, and data-centric augmentation rather than a single standardized procedure.

Imbalance (15.5% = 13/84 studies; e.g., IDs 9, 30, 36): This category combines approaches for handling skewed correctness distributions with data-centric augmentation. The problem occurs when certain skills are predominantly answered correctly (very easy) or incorrectly (very difficult), driving biased mastery estimates. Reported mitigations include: (1) contrastive or re-weighted representation learning modules that explicitly balance local and global interaction frequencies to keep minority trajectories salient (e.g., IDs 74, 75); (2) preprocessing heuristics such as filtering low-activity learners, truncating extreme frequencies, or normalizing heterogeneous features before model fitting (e.g., IDs 30, 44, 57); (3) imbalance-aware embeddings that adapt student–concept representations when population sizes differ markedly (e.g., ID 55); (4) relational propagation that leverages graph structure to smooth sparse curriculum regions and reduce majority-class dominance (e.g., ID 69); and (5) stratified sampling and autoencoder-based augmentation that manufacture additional minority interaction sequences (e.g., ID 80). Several studies simply acknowledge imbalance and rely on robust metrics such as AUC or balanced accuracy instead of explicit treatments (e.g., IDs 9, 38, 49, 51, 68), underscoring the field's ongoing reliance on evaluation-side workarounds rather than holistic data rebalancing.

Data noise (1.2% = 1/84 studies; e.g., ID 82): Although rare, some work explicitly tackles logging artefacts and duplicated entries that distort learner histories. The representative study refreshes the

ASSISTments release to merge duplicate interactions, discards incomplete records, and layers dropout regularization to suppress residual noise. These preprocessing pipelines emphasize the importance of curating interaction traces before model training when raw exports contain inconsistent or partially recorded events.

Q-Matrix bias (3.6% = 3/84 studies; e.g., IDs 3, 64, 72): Expert-defined Q-matrices, which specify the skills required for each question, often contain subjective biases or may be incomplete. To address this, studies have applied: (1) learnable skill embeddings, enabling the model to automatically discover skill–question relationships and override expert-defined mappings when they conflict with observed data (e.g., IDs 3, 64, 72); (2) automatic Q-matrix expansion, which identifies additional skill–question relationships not captured in the original expert specification (e.g., ID 64); and (3) skill–question relationship refinement, which adjusts the strength of expert-defined links based on empirical evidence (e.g., ID 72).

Causal de-bias (1.2% = 1/84 studies; e.g., ID 62): This emerging area focuses on algorithmic fairness and the mitigation of spurious correlations in knowledge tracing. The representative study integrates front-door adjustments into a causal self-attention mechanism to dampen confounding factors and recover more equitable mastery trajectories across demographic groups. Although nascent, this causal treatment illustrates how fairness-aware objectives can be embedded directly in the representation learning pipeline when data bias poses risks to student-facing predictions.

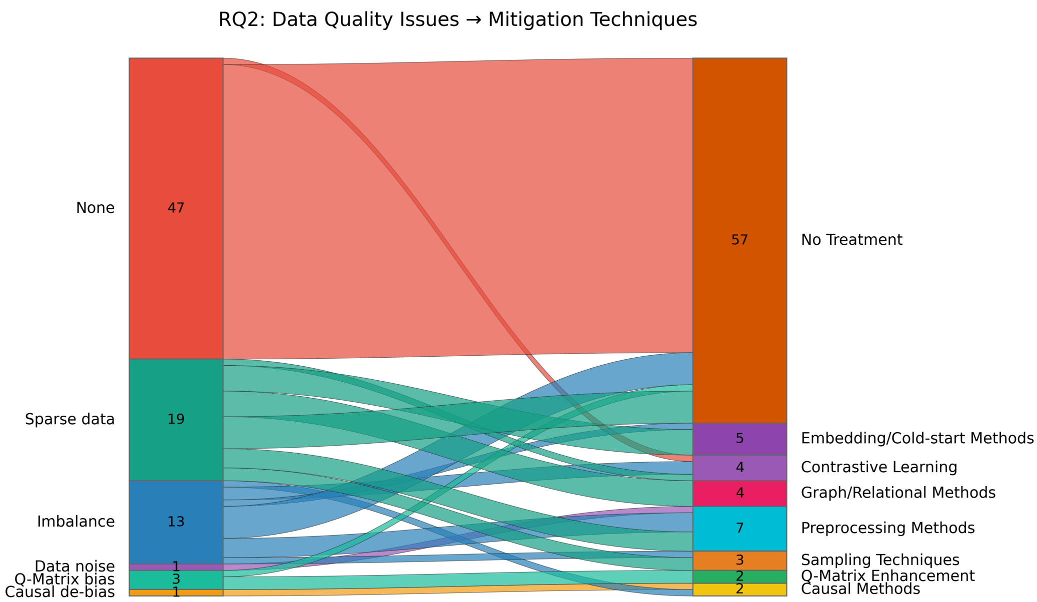

Figure 6. Parallel sets diagram showing the flow from data quality issues (left) to their treatment techniques (right) in DKT research. Node sizes represent the number of papers addressing each issue, with flowing connections indicating which techniques are used for each issue type.

3.3. RQ3: Sequential Stability

Sequential stability refers to a model's ability to generate smooth trajectories of a learner's latent knowledge state as additional interactions are observed. This concept is important for three reasons. First, rapid oscillations in predicted mastery undermine the practical value of educational AI systems: teachers and students cannot rely on feedback that reverses direction after a single response, especially in contexts like K–12 adaptive tutoring where real-time instructional decisions are needed. Second, instabilities often indicate overfitting to unusual sequences and therefore predict poor generalization to unseen learners. Third, the cost of misclassification can vary across different points in a learning sequence: for example, mistakenly predicting mastery early on may prevent crucial practice opportunities, while a similar error near the end of instruction might carry less consequence, mainly influencing final assessment rather than ongoing learning. Despite its importance, sequential stability

 doi:10.20944/preprints202510.1845.v1

has received far less attention compared to commonly reported accuracy metrics such as AUC or RMSE.

Regarding sequential stability of predictions, our synthesis (Table 4) shows that *qualitative visual inspection* is by far the predominant method. Among the investigations reviewed, 62 of 84 (73.8%) present exemplar visualizations—typically heat maps of item-level probabilities or line charts of mastery trajectories—for a small number of learners. These graphical representations support model intuition but are typically applied to only a few test cases and are rarely paired with objective assessment criteria, offering limited evidence of reliability across larger learner populations. A smaller subset—3 of 84 (3.6%; IDs 29, 30, 77)—computes *quantitative stability metrics*, including the *prediction-waviness score* (sum of absolute first-order differences across trajectories), ℓ_2-norm fluctuations between successive knowledge vectors, and the *temporal-consistency coefficient* originally introduced by Piech et al. [11]. These metrics capture different aspects of stability: waviness reflects smoothness, while temporal consistency reflects predictive coherence across time. Such measures enable direct model comparison but remain under-reported and lack established thresholds for acceptable stability [14].

Approximately 6.0% of studies (5 of 84) examine forgetting behavior, testing whether predicted proficiency decays during inactivity. This is useful for detecting models that unrealistically preserve full mastery indefinitely, though most analyses rely on synthetic rather than authentic gaps (e.g., IDs 2, 16, 19, 65). Only 1 paper extends evaluation to perturbation-based sensitivity analysis (1.2% of 84), injecting random noise into input sequences and measuring the resulting trajectory variance (ID 80).

Overall, no study in this corpus jointly reports stability metrics, forgetting, and sensitivity, indicating that evaluations of sequential stability remain fragmented rather than holistic.

Table 4. Evaluation of sequential stability in empirical modeling studies.

Sequential stability approach	**Papers**	**%**	**Typical techniques**	**Paper IDs**
Not evaluated	13	15.5%	—	1, 8, 11, 18, 21, 35, 36, 37, 44, 51, 55, 73, 76
Qualitative visualization / examples	62	73.8%	Heat-maps or illustrative mastery-trajectory plots	3, 4, 5, 6, 7, 9, 10, 12, 13, 14, 17, 20, 22, 23, 24, 25, 26, 27, 28, 31, 32, 33, 34, 38, 39, 40, 41, 42, 43, 45, 46, 47, 48, 49, 50, 52, 53, 54, 56, 57, 58, 59, 60, 61, 62, 63, 64, 66, 67, 68, 69, 70, 71, 72, 74, 75, 78, 79, 81, 82, 83, 84
Quantitative stability metrics	3	3.6%	Prediction waviness, ℓ_2-norm fluctuations, temporal-consistency coefficient	29, 30, 77
Forgetting / decay analysis	5	6.0%	Forgetting-curve tracking and decay-aware proficiency plots	2, 15, 16, 19, 65
Sensitivity / robustness analysis	1	1.2%	Perturbation-based sensitivity analysis	80

†Counts are not mutually exclusive; a single study may employ more than one evaluation.

3.4. RQ4: Process-Level Interpretability

Process-level interpretability refers to approaches that reveal how a knowledge tracing model generates predictions across a learning sequence, rather than focusing solely on end-point accuracy. In DKT, this entails clarifying the mechanisms, structures, or inputs that drive the evolution of a learner's latent knowledge state and its observable outputs.

As summarized in Table 5, process-level interpretability remains underexplored in contemporary DKT research. A majority of studies (88.1%, 74 of 84) do not consider interpretability at all, offering

 doi:10.20944/preprints202510.1845.v1

no explicit strategies for explaining how predictions are derived. Among the remaining works that incorporate process-level interpretability, the most frequently implemented approach is feature attribution / attention analysis (9.5% of studies; e.g., IDs 1, 11, 31). These post-hoc techniques seek to explain a model's current prediction by isolating which prior learner interactions exerted the greatest influence. Approaches include Deep SHAP (ID 11) for Shapley-based contribution scoring of historical interactions, attention-weight saliency inspection (e.g., IDs 31, 71) to highlight pedagogically salient exercises emphasized by Transformer or hierarchical attentive modules, and sequence-adapted gradient propagation methods such as Grad-CAM (ID 71) to spatially (temporally) localize decisive segments within interaction histories. An example of analytic feature saliency based on model sensitivity for this purpose was given in Linja et al. [70]. While these techniques provide granular, instance-level justifications, their interpretive value for educators is moderated by two limitations: (i) attention distributions or attribution scores do not always correspond to causal pedagogical factors, risking over-interpretation; and (ii) their effective use often presupposes technical literacy to contextualize attributions within curriculum design or assessment strategies.

Table 5. Interpretability techniques reported in empirical DKT studies.

Interpretability approach	Papers	%	Typical techniques	Paper IDs
Not considered	74	88.1%	—	2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 14, 15, 16, 17, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 63, 64, 65, 66, 67, 68, 69, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84
Interpretable architecture (IRT / rule-based layer)	1	1.2%	IRT prediction layer; concept-level factorisation; rule-based gating	60
Feature attribution / attention analysis	8	9.5%	SHAP, attention-weight analyses, Grad-CAM visualisations	1, 8, 11, 19, 21, 31, 70, 71
Embedding or trajectory visualisation	1	1.2%	t-SNE plots of knowledge embeddings to reveal skill clusters	46

†Counts are not mutually exclusive; a single study may employ more than one interpretability approach.

The second most common architectural strategy is interpretability by design (1.2% of all studies; e.g., ID 60). These models embed educationally meaningful inductive structure directly into the architecture—such as skill- or concept-level parameterisation / factorisation layers or decomposable outputs—so intermediate variables can be audited. Such structural constraints increase semantic alignment with instructional constructs and can surface intermediate variables that are straightforward to audit. However, they often operate primarily as knowledge injection to enhance predictive calibration rather than fully disclosing internal state transitions or uncertainty propagation. Thus, transparency is partial: educators gain visibility into selected pedagogically grounded components but not necessarily into the full interaction of latent subsystems.

Finally, a small subset investigates embedding or trajectory visualization (1.2% of studies; ID 46), applying methods such as t-SNE projections of latent knowledge embeddings to reveal emergent clustering of skills or learner progression pathways. These visual artefacts offer intuitive, exploratory overviews of representational geometry but remain qualitative, lacking explicit alignment to decision-relevant thresholds or actionable formative feedback. They therefore complement but do not substitute for attribution- or design-oriented interpretability.

4. Discussion

This systematic review highlights notable advances in predictive learner modeling while also surfacing new insights into unresolved challenges that hinder the responsible deployment of DKT models in practice. In particular, it reveals a narrow modeling landscape dominated by a single dataset family, limited attention to data quality, and the near absence of sequential stability assessments. It also uncovers a critical transparency gap, as interpretability remains underdeveloped despite growing educational and regulatory demands. These findings together provide a novel, cross-cutting perspective on the systemic challenges that must be addressed for DKT to achieve sustainable impact.

4.1. Modeling Landscape and Architectural Evolution

The temporal evolution patterns of DKT architectures reveal a field that has progressed from early sequence modeling approaches to a diverse ecosystem dominated by Graph-Based models (26.2% of studies), followed by Hybrid/Meta (23.8%) and Attentive models (17.9%). This trajectory illustrates the field's shift from simple sequential models toward architectures that promise richer relational reasoning and adaptive flexibility.

Dataset utilization patterns reveal a concerning over-reliance on ASSIST datasets, employed by 82.1% of studies. While these datasets provide valuable benchmarks, this concentration may perpetuate methodological blind spots and limit the generalizability of findings across diverse educational contexts [77]. The field's implicit trust in established benchmarks potentially masks data quality issues that become apparent only when models encounter varied educational environments, as demonstrated by the finding that studies using diverse datasets show similar rates of data quality consideration (47%) compared to those relying solely on ASSIST (38%) [78].

Regarding the modeled characteristics, there is a notable trend towards incorporating more complex skill representations and relationships. Many studies (e.g., IDs 4, 6) leverage graph-based structures to model the interdependencies between skills, allowing for a more nuanced understanding of learner trajectories. This shift is indicative of a broader recognition of the importance of relational reasoning in educational contexts. Concerning the evaluation metrics employed for model assessment, the dominance of AUC (reported in 90.5% of studies) demonstrates near-universal adoption of a single discriminative benchmark. However, this heavy reliance reinforces a predominantly binary predictive framing that neglects sequential dependencies and temporal dynamics central to learning processes.

The implications of this architectural and geographical concentration extend beyond academic concerns. The dominance of certain research traditions and dataset preferences may inadvertently embed cultural and pedagogical assumptions into model architectures, potentially limiting their effectiveness when deployed in different educational contexts [60,77]. Future research must prioritize architectural designs that account for diverse learning environments and cultural contexts to ensure global applicability (Chinta et al. [79]).

4.2. Data Quality and Methodological Rigor

The critical evaluation of data quality treatment reveals one of the most concerning findings of this review: 56.0% of studies fail to acknowledge or address fundamental data quality issues inherent in educational datasets. This methodological limitation raises important concerns for the real-world deployment and educational validity of DKT systems, as it may contribute to biased assessments and reduce the reliability of recommendations [80,81].

The categorization of bias mitigation strategies demonstrates that when studies do address data quality, they predominantly focus on sparse data (22.6% of studies) and imbalance (15.5% of studies). These approaches often rely on architectural innovations rather than systematic data preprocessing, reflecting the sequential nature of knowledge tracing data where traditional resampling techniques risk disrupting temporal dependencies fundamental to accurate prediction [82].

The relationship between model architecture and data quality consideration shows a qualitative pattern: models that explicitly encode relational or structural information (e.g., graph-based or memory-augmented designs) more frequently describe data validation or augmentation steps than purely sequential baselines, likely because structural priors make inconsistencies in mappings and sparsity more visible [78,83].

The lack of systematic approaches to imbalance is particularly problematic given recent findings that mainstream knowledge tracing models show significant performance drops when evaluated on resampled, balanced test sets, revealing their heavy reliance on answer distribution biases rather than genuine learning patterns [56]. This suggests that the predictive accuracy reported in much of the existing literature may overestimate real-world performance and educational utility [30]. The emerging attention to causal de-bias (1.2% of studies) represents a critical but underexplored area. As educational AI systems directly impact students' academic trajectories and future opportunities, ensuring algorithmic fairness across demographic groups becomes not merely a technical concern but an ethical imperative [18,79]. The field must urgently develop standardized protocols for bias detection and mitigation that account for the unique characteristics of educational data while ensuring equitable outcomes across diverse student populations [14,81].

4.3. Sequential Stability Assessment

The systematic evaluation of sequential stability of predictions reveals one of the most significant unresolved problem in current DKT research. Only 3.6% of studies employ quantitative sequential stability metrics, despite the critical importance of stable, educationally plausible knowledge trajectories for practical deployment in educational settings [44,45].

Variation across architectural paradigms is described but rarely systematically quantified in the source studies. Designs that explicitly encode temporal decay or memory structures more often report qualitative trajectory analyses, whereas pure attention-based architectures seldom introduce additional stability diagnostics beyond illustrative plots [62,78]. The implications of sequential instability extend beyond theoretical concerns. In practical educational settings, teachers and adaptive learning systems require models that produce stable, interpretable progressions of student knowledge. A model that achieves high predictive accuracy but exhibits "wavy transitions"—predicting mastery after one correct answer only to reverse this assessment after the next—provides little actionable guidance and may undermine educator trust in AI-driven insights [61,63].

The field's evolution toward responsible AI in education necessitates fundamental reconsideration of evaluation practices. Future work must prioritize metrics that explicitly account for sequential dependencies and educational validity, including trajectory smoothness indices, learning progression coherence measures, and temporal consistency evaluations that ensure predicted knowledge sequences follow established developmental pathways [84,85].

4.4. Interpretability and Transparency

The analysis of interpretability mechanisms reveals a critical transparency deficit that directly conflicts with emerging regulatory frameworks and educational needs. Only 11.9% of studies provide any form of interpretability, with 88.1% deploying black-box architectures that offer no insight into their decision-making processes [59,64]. Model complexity shows an inverse relationship with interpretability. Memory-centric or factorised representations can support state inspection, whereas heterogeneous hybrid architectures rarely embed coherent explanatory interfaces [78]. This reflects a broader trade-off: as architectures grow more sophisticated to improve predictive accuracy, their internal reasoning often becomes more opaque [78].

In educational contexts, the distinction between interpretability and explainability is critical [67,68,86]. Most studies that attempt prediction explanation rely on post-hoc methods, such as SHAP or Grad-CAM, which may not faithfully capture the true decision-making process [16]. By contrast, interpretability-by-design approaches constrain models to remain transparent but may sacrifice predictive power [67,68,86]. This opacity creates both educational and regulatory risks. Teachers cannot

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2025 doi:10.20944/preprints202510.1845.v1

meaningfully integrate AI-driven insights into practice without understanding the reasoning behind mastery predictions [59,64]. Under the European Union's AI Act, deploying black-box models for student assessment may become legally non-compliant, creating both practical limitations and regulatory vulnerabilities for 88.1% of current models [87].

4.5. Cross-Cutting Patterns Across Responsible AI Dimensions

The synthesis across the three responsible AI dimensions—data quality, sequential stability, and interpretability—shown in Figure 7, reveals clear, category-specific patterns with direct implications for responsible deployment in education [14,18].

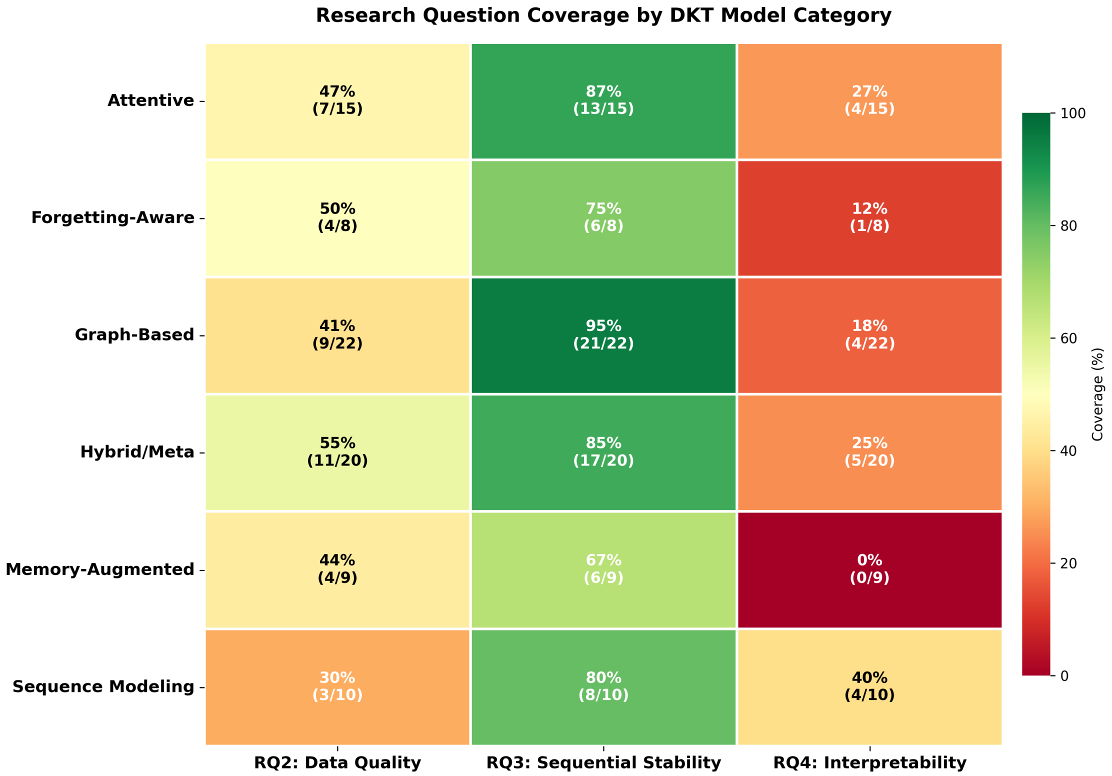

Figure 7. Research Question coverage by model category showing the percentage of papers in each category that address RQ2 (Data Quality), RQ3 (Sequential Stability), and RQ4 (Interpretability).

The first tension involves the conflict between predictive performance optimization and educational validity. Studies reporting the highest AUC scores (>0.85) rarely provide complementary interpretability or stability evidence, amplifying a performance–validity gap [72,88]. The field's heavy reliance on AUC as the dominant metric (90.5% of studies) assumes independence of observations, a mismatch that risks incentivizing models which exploit statistical regularities rather than capturing authentic learning dynamics [89,90].

The second tension arises from dataset concentration effects. Over-reliance on ASSIST datasets (82.1% of studies) fosters implicit trust in established benchmarks while potentially obscuring data quality issues. In contrast, studies using more diverse datasets more frequently address methodological concerns across multiple dimensions, suggesting that benchmark diversity can function as a form of natural quality assurance [77,78].

4.6. Limitations

This review has inherent limitations that must be acknowledged. The English-only inclusion criterion may have excluded relevant research from non-English speaking countries. The taxonomic application required some subjective judgment despite clear criteria, and quantitative meta-analysis was not feasible due to heterogeneous reporting standards. While three researchers participated in this process to reduce bias, the classification may not be entirely free from subjectivity.

Our search strategy, though comprehensive across five major databases and grey literature, may also carry selection biases. The keyword combinations centered on "knowledge tracing" with ("deep*" OR "neural*") may have excluded studies using alternative terms such as "student modeling." Upon extensive meta-review of the existing works, we decided that DKT studies would reference these terms in their title, abstract, or keywords, though it remains possible that a very few studies might have been missed. The grey literature search via Google Scholar helped to mitigate this risk, although expanding the search further could potentially have identified additional relevant materials. Additionally, while this review covers diverse architectures including transformers, graph neural networks, and memory-augmented models, the original formulation of DKT focused on recurrent architectures, which shaped early field development and may have influenced the conceptual framing of knowledge tracing as primarily a recurrent modeling task. Finally, the review's focus on empirical modeling studies means that theoretical contributions and methodological frameworks not directly tied to model development may be underrepresented.

5. Conclusions

This systematic review identifies both significant achievements and unresolved problems in DKT research from 2015–2025. While predictive accuracy has improved substantially through architectural innovations, the analysis reveals persistent problems that prevent practical deployment and compromise the educational validity of these models.

The most concerning finding is the widespread neglect of fundamental methodological requirements essential for responsible AI deployment in education. Only 3.6% of studies employ quantitative sequential stability assessment, despite its critical importance for generating educationally plausible knowledge trajectories. A majority (56.0%) fail to address data quality issues inherent in educational datasets, potentially compromising model reliability and fairness. Furthermore, 88.1% lack interpretability mechanisms, directly conflicting with emerging regulatory frameworks such as the EU AI Act, which mandates transparency for high-risk AI applications in education with full enforcement by August 2026.

The field's over-reliance on AUC as the primary evaluation metric (90.5% of studies) reflects a fundamental methodological mismatch with the sequential nature of learning data. This emphasis on aggregate performance metrics may inadvertently incentivize models that exploit statistical regularities rather than capture genuine learning dynamics, undermining educational utility despite high reported accuracy. It also reinforces a binary framing of student knowledge, reducing inherently continuous and latent learning processes to discrete correctness outcomes. The geographic concentration of research and overwhelming dependence on ASSIST datasets (82.1% of studies) create potential blind spots that may limit the global applicability and cultural sensitivity of DKT systems. Studies utilizing more diverse datasets demonstrate higher rates of methodological rigor across multiple dimensions, suggesting that benchmark diversity can serve as a natural quality assurance mechanism. The synthesis reveals that methodological advancement has not kept pace with architectural innovation. Only a very small minority of studies address all four critical dimensions (modeling, data quality, sequential stability, and interpretability) with substantive depth. No single architectural category demonstrates consistent methodological strength across all dimensions, challenging the assumption that computational sophistication naturally leads to educational validity.

As AI systems become increasingly integrated into educational settings, the need for responsible AI frameworks encompassing transparency, fairness, accountability, and educational validity becomes paramount. The field has to evolve beyond accuracy-focused development toward educationally valid, stable, and transparent models that serve educational stakeholders effectively. This evolution is not merely academically desirable but increasingly legally necessary, as emerging regulatory frameworks impose mandatory transparency requirements for educational AI systems.

 doi:10.20944/preprints202510.1845.v1

Future research should prioritize standardized protocols for evaluating sequential stability, systematic methods for addressing imbalance, and inherently interpretable architectures that embed transparency into core design. Expanding validation beyond common datasets to diverse educational contexts is essential for ensuring global applicability. DKT research should also integrate insights from cognitive science and learning theory to ensure that advances in modeling translate into meaningful educational outcomes.

Acknowledgments: This work was supported by the Estonian Research Council grant (PRG2215).

Appendix A. Reference Table with Key Characteristics

Table A1 presents the complete list of 84 screened papers with their bibliographic information and key characteristics extracted during the systematic review process. Each row represents one study, with columns providing structured data for systematic comparison across the corpus.

The table columns capture the following dimensions analyzed in this review. **ID** provides a unique identifier number assigned to each paper for cross-referencing throughout the review. **Reference** contains the complete bibliographic citation following standard academic format. **Category** indicates the primary architectural classification based on the adapted taxonomy used in this review, including Attentive, Graph-Based, Sequence Modeling, Memory-Augmented, Forgetting-Aware, and Hybrid/Meta. Models incorporating textual features are mapped to these categories according to their dominant computational mechanism (e.g., attention-driven text models are categorized as Attentive). **Dataset** specifies the educational datasets used for model evaluation, such as ASSIST, EdNet, Statics2011, and Programming datasets. **Metrics** lists the evaluation metrics employed to assess model performance, including AUC, ACC, and RMSE. **Data Quality** describes the approach to addressing data quality issues, with values including "None" for no explicit treatment, "Sparse data" for handling missing interactions, "Imbalance" for class imbalance mitigation, "Data noise" for noise filtering, "Q-Matrix bias" for expert bias correction, and "Causal de-bias" for fairness considerations. **Stability** indicates the type of sequential stability evaluation performed, ranging from "None" for not evaluated, "Visual" for qualitative visualization, and "Quantitative" for numerical stability metrics, "Forgetting" for decay analysis, to "Sensitivity" for robustness analysis. **Interpretability** specifies the mechanism for model explainability, with values aligned to Table 5: "Not considered" for studies without interpretability, "Feature attribution / attention analysis" for post-hoc attributions, "Embedding or trajectory visualisation" for latent-space diagnostics, and "Interpretable architecture (IRT / rule-based layer)" for by-design transparency.

doi:10.20944/preprints202510.1845.v1

Table A1. Reference table with key characteristics of screened papers.

ID	Reference	Category	Dataset	Metrics	Data Quality	Stability	Interpretability
1	Li, et al. (2024). A Genetic Causal Explainer for Deep Knowledge Tracing.	Hybrid/Meta	ASSIST09, EdNet	AUC	None	None	Feature attribution / attention analysis
2	Cheng, et al. (2022). A Knowledge Query Network Model Based on Rasch Model Embedding for Personalized Online Learning.	Attentive	ASSIST09, ASSIST15, ASSIST17	AUC	None	Forgetting	Not considered
3	Zhao, et al. (2023). A novel framework for deep knowledge tracing via gating-controlled forgetting and learning mechanisms.	Attentive	ASSIST15, ASSIST17, Junyi	AUC, ACC	Q-Matrix bias	Visual	Not considered
4	Liu, et al. (2024). A probabilistic generative model for tracking multi-knowledge concept mastery probability.	Graph-Based	Algebra2006, Algebra2008, POJ	AUC, ACC, RMSE, MAE, MSE, Precision, Recall	Sparse data	Visual	Not considered
5	Zhang, et al. (2024). A Question-centric Multi-experts Contrastive Learning Framework for Improving the Accuracy and Interpretability of Deep Sequential Knowledge Tracing Models.	Attentive	ASSIST09, EdNet, Algebra2005	AUC, ACC	None	Visual	Not considered
6	Liu, et al. (2022). Ability boosted knowledge tracing.	Graph-Based	ASSIST09, AICFE	AUC, ACC	None	Visual	Not considered
7	Cheng, et al. (2022). AdaptKT: A Domain Adaptable Method for Knowledge Tracing.	Attentive	ASSIST09, ASSIST15, ASSIST17	AUC, F1	Sparse data	Visual	Not considered
8	Wang, et al. (2023). An Efficient and Generic Method for Interpreting Deep Learning based Knowledge Tracing Models.	Sequence Modeling	ASSIST09, ASSIST15	AUC	None	None	Feature attribution / attention analysis
9	Luo, et al. (2024). An efficient state-aware Coarse-Fine-Grained model for Knowledge Tracing.	Attentive	ASSIST15, ASSIST17, Junyi	AUC, ACC, RMSE, F1	Imbalance	Visual	Not considered
10	Shen, et al. (2022). Assessing Student's Dynamic Knowledge State by Exploring the Question Difficulty Effect.	Hybrid/Meta	ASSIST12, Eedi	AUC, ACC, RMSE	None	Visual	Not considered

 doi:10.20944/preprints202510.1845.v1

Table A1. *Cont.*

ID	Reference	Category	Dataset	Metrics	Data Quality	Stability	Interpretability
11	Xu, et al. (2024). Bridging the Vocabulary Gap: Using Side Information for Deep Knowledge Tracing.	Attentive	Private dataset	AUC	Sparse data	None	Feature attribution / attention analysis
12	Zu, et al. (2023). CAKT: Coupling contrastive learning with attention networks for interpretable knowledge tracing.	Attentive	ASSIST09, ASSIST15, ASSIST17	AUC	None	Visual	Not considered
13	Li, et al. (2023). Calibrated Q-Matrix-Enhanced Deep Knowledge Tracing with Relational Attention Mechanism.	Graph-Based	ASSIST12, Eedi	AUC, ACC	None	Visual	Not considered
14	Ghosh, et al. (2020). Context-Aware Attentive Knowledge Tracing.	Attentive	ASSIST09, ASSIST15, ASSIST17	AUC	None	Visual	Not considered
15	Chen, et al. (2022). DCKT: A Novel Dual-Centric Learning Model for Knowledge Tracing.	Attentive	ASSIST09, ASSIST12, ASSIST15	AUC	None	Forgetting	Not considered
16	Abdelrahman, et al. (2023). Deep Graph Memory Networks for Forgetting-Robust Knowledge Tracing.	Forgetting-Aware	ASSIST09, Statics2011, KDD2010	AUC, ACC, Loss	None	Forgetting	Not considered
17	Tsutsumi, et al. (2024). Deep Knowledge Tracing Incorporating a Hypernetwork With Independent Student and Item Networks.	Forgetting-Aware	ASSIST09, ASSIST15, ASSIST17	AUC, ACC, Loss	None	Visual	Not considered
18	Yang, et al. (2022). Deep Knowledge Tracing with Learning Curves.	Attentive	ASSIST09, ASSIST15, ASSIST17	AUC	Sparse data	None	Not considered
19	Wang, et al. (2024). Deep Knowledge Tracking Integrating Programming Exercise Difficulty and Forgetting Factors.	Hybrid/Meta	Other dataset	AUC, F1	None	Forgetting	Feature attribution / attention analysis
20	Tsutsumi, et al. (2024). Deep-IRT with a Temporal Convolutional Network for Reflecting Students' Long-Term History of Ability Data.	Sequence Modeling	ASSIST09, ASSIST17, Junyi	AUC, ACC	None	Visual	Not considered

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2025 doi:10.20944/preprints202510.1845.v1

Table A1. *Cont.*

ID	Reference	Category	Dataset	Metrics	Data Quality	Stability	Interpretability
21	Lu, et al. (2024). Design and evaluation of Trustworthy Knowledge Tracing Model for Intelligent Tutoring System.	Hybrid/Meta	ASSIST09	AUC	None	None	Feature attribution / attention analysis
22	Wang, et al. (2024). DF-EGM: Personalised Knowledge Tracing with Dynamic Forgetting and Enhanced Gain Mechanisms.	Hybrid/Meta	ASSIST09, ASSIST12, Algebra2005	AUC, ACC	None	Visual	Not considered
23	Kim, et al. (2021). DiKT: Dichotomous Knowledge Tracing.	Sequence Modeling	ASSIST09, ASSIST15, Statics2011	AUC, ACC	None	Visual	Not considered
24	Zhou, et al. (2024). Discovering Multi-Relational Integration for Knowledge Tracing with Retentive Networks.	Forgetting-Aware	ASSIST09, ASSIST12, EdNet	AUC	None	Visual	Not considered
25	Zhang, et al. (2024). DKVMN-KAPS: Dynamic Key-Value Memory Networks Knowledge Tracing With Students' Knowledge-Absorption Ability and Problem-Solving Ability.	Memory-Augmented	ASSIST09, ASSIST15, Statics2011	AUC	None	Visual	Not considered
26	Xu, et al. (2024). DKVMN&MRI: A new deep knowledge tracing model based on DKVMN incorporating multi-relational information.	Graph-Based	ASSIST09, ASSIST15, EdNet	AUC, ACC	None	Visual	Not considered
27	Wang, et al. (2023). Dynamic Cognitive Diagnosis: An Educational Priors-Enhanced Deep Knowledge Tracing Perspective.	Hybrid/Meta	ASSIST09, ASSIST12	AUC, ACC	Sparse data	Visual	Not considered
28	Liu, et al. (2019). EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction.	Sequence Modeling	Other dataset	AUC	Sparse data	Visual	Not considered
29	Pu, et al. (2024). ELAKT: Enhancing Locality for Attentive Knowledge Tracing.	Graph-Based	Private dataset	AUC, ACC, RMSE, MAE	None	Quantitative	Not considered
30	Cai, et al. (2025). Enhanced Knowledge Tracing via Frequency Integration and Order Sensitivity.	Forgetting-Aware	ASSIST09, ASSIST17, Statics2011	AUC, ACC, MAE	Imbalance	Quantitative	Not considered

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2025 doi:10.20944/preprints202510.1845.v1

Table A1. *Cont.*

ID	Reference	Category	Dataset	Metrics	Data Quality	Stability	Interpretability
31	Qian, et al. (2024). Enhanced Knowledge Tracing With Learnable Filter.	Hybrid/Meta	ASSIST09, ASSIST12, ASSIST15	AUC	None	Visual	Feature attribution / attention analysis
32	Liu, et al. (2023). Enhancing Deep Knowledge Tracing with Auxiliary Tasks.	Hybrid/Meta	Algebra2005, Algebra2006, Bridge	AUC, ACC	None	Visual	Not considered
33	Guo, et al. (2021). Enhancing Knowledge Tracing via Adversarial Training.	Graph-Based	ASSIST09, ASSIST15, ASSIST17	AUC	None	Visual	Not considered
34	Zhao, et al. (2023). Exploiting multiple question factors for knowledge tracing.	Hybrid/Meta	ASSIST09, EdNet, KT1	AUC, ACC	Sparse data	Visual	Not considered
35	Zhang, et al. (2024). Explore Bayesian analysis in Cognitive-aware Key-Value Memory Networks for knowledge tracing in online learning.	Memory-Augmented	ASSIST09, ASSIST17, Algebra2005	AUC	Sparse data	None	Not considered
36	Wu, et al. (2021). Federated Deep Knowledge Tracing.	Sequence Modeling	ASSIST (unspecified), MATH	AUC, ACC, RMSE, MSE	Imbalance	None	Not considered
37	Hooshyar, et al. (2022). GameDKT: Deep knowledge tracing in educational games.	Hybrid/Meta	Taiwan	AUC, ACC, Precision, Recall	Sparse data	None	Not considered
38	Liang, et al. (2024). GELT: A graph embeddings based lite-transformer for knowledge tracing.	Attentive	ASSIST09, ASSIST12, EdNet	AUC	Imbalance	Visual	Not considered
39	Yanyou, et al. (2024). Global Feature-guided Knowledge Tracing.	Hybrid/Meta	ASSIST12, ASSIST17, FrcSub	AUC, ACC, RMSE	None	Visual	Not considered
40	Nakagawa, et al. (2019). Graph-based knowledge tracing: Modeling student proficiency using graph neural networks.	Graph-Based	ASSIST09	AUC	None	Visual	Not considered
41	Wang, et al. (2023). GraphCA: Learning from Graph Counterfactual Augmentation for Knowledge Tracing.	Graph-Based	ASSIST09, ASSIST12, Algebra2006	AUC, ACC	Sparse data	Visual	Not considered
42	Liu, et al. (2024). Heterogeneous Evolution Network Embedding with Temporal Extension for Intelligent Tutoring Systems.	Graph-Based	ASSIST09, ASSIST12, EdNet	AUC, ACC	None	Visual	Not considered

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2025 doi:10.20944/preprints202510.1845.v1

Table A1. *Cont.*

ID	Reference	Category	Dataset	Metrics	Data Quality	Stability	Interpretability
43	Yang, et al. (2024). Heterogeneous graph-based knowledge tracing with spatiotemporal evolution.	Graph-Based	Private dataset	AUC, ACC	None	Visual	Not considered
44	Yang, et al. (2018). Implicit Heterogeneous Features Embedding in Deep Knowledge Tracing.	Memory-Augmented	ASSIST (unspecified), Junyi	AUC	Imbalance	None	Not considered
45	Chen, et al. (2023). Improving Interpretability of Deep Sequential Knowledge Tracing Models with Question-centric Cognitive Representations.	Graph-Based	ASSIST09, Algebra2005	AUC, ACC	None	Visual	Not considered
46	Xu, et al. (2023). Improving knowledge tracing via a heterogeneous information network enhanced by student interactions.	Graph-Based	ASSIST09, EdNet, Statics2011	AUC, ACC	None	Visual	Embedding or trajectory visualisation
47	Shu, et al. (2024). Improving Knowledge Tracing via Considering Students' Interaction Patterns.	Memory-Augmented	ASSIST09, ASSIST12, ASSIST17	AUC, ACC	None	Visual	Not considered
48	Long, et al. (2022). Improving Knowledge Tracing with Collaborative Information.	Memory-Augmented	EdNet	AUC, ACC	None	Visual	Not considered
49	Tato, et al. (2022). Infusing Expert Knowledge Into a Deep Neural Network Using Attention Mechanism for Personalized Learning Environments.	Graph-Based	Other dataset	ACC, F1, Precision, Recall	Imbalance	Visual	Not considered
50	Zhang, et al. (2021). Input-Aware Neural Knowledge Tracing Machine.	Graph-Based	ASSIST09, ASSIST12, Algebra2006	AUC, ACC	None	Visual	Not considered
51	He, et al. (2023). Integrating fine-grained attention into multi-task learning for knowledge tracing.	Forgetting-Aware	ASSIST09, ASSIST17, Statics2011	AUC, ACC	Imbalance	None	Not considered
52	Chen, et al. (2024). Interaction Sequence Temporal Convolutional Based Knowledge Tracing.	Sequence Modeling	ASSIST09, ASSIST12, Algebra2005	AUC, ACC	None	Visual	Not considered
53	Sun, et al. (2024). Interpretable Knowledge Tracing with Multiscale State Representation.	Sequence Modeling	ASSIST09, ASSIST12, EdNet	AUC	Sparse data	Visual	Not considered

doi:10.20944/preprints202510.1845.v1

Table A1. *Cont.*

ID	Reference	Category	Dataset	Metrics	Data Quality	Stability	Interpretability
54	Tong, et al. (2022). Introducing Problem Schema with Hierarchical Exercise Graph for Knowledge Tracing.	Graph-Based	Other dataset	AUC, ACC	None	Visual	Not considered
55	Song, et al. (2021). JKT: A joint graph convolutional network based Deep Knowledge Tracing.	Graph-Based	ASSIST09, ASSIST15, ASSISTChall	AUC, ACC	Imbalance	None	Not considered
56	Pan, et al. (2024). Knowledge Graph and Personalized Answer Sequences for Programming Knowledge Tracing.	Attentive	Other dataset	AUC, ACC	None	Visual	Not considered
57	Gan, et al. (2022). Knowledge interaction enhanced sequential modeling for interpretable learner knowledge diagnosis in intelligent tutoring systems.	Hybrid/Meta	Statics2011, Algebra2005, Algebra2006	AUC, ACC, Loss	Imbalance	Visual	Not considered
58	Lee & Yeung (2019). Knowledge Query Network for Knowledge Tracing.	Graph-Based	ASSIST09, ASSIST15, Statics2011	AUC	None	Visual	Not considered
59	Gan, et al. (2022). Knowledge structure enhanced graph representation learning model for attentive knowledge tracing.	Graph-Based	ASSIST09, ASSIST12, EdNet	AUC	Sparse data	Visual	Not considered
60	Xiao, et al. (2023). Knowledge tracing based on multi-feature fusion.	Sequence Modeling	ASSIST09, ASSIST15, Statics2011	AUC	None	Visual	Interpretable architecture (IRT / rule-based layer)
61	Xu, et al. (2023). Learning Behavior-oriented Knowledge Tracing.	Attentive	ASSIST09, ASSIST12, Junyi	AUC, ACC, RMSE	None	Visual	Not considered
62	Huang, et al. (2024). Learning consistent representations with temporal and causal enhancement for knowledge tracing.	Hybrid/Meta	ASSIST12, EdNet, KT1	AUC, ACC, RMSE	Causal de-bias	Visual	Not considered
63	Wang, et al. (2021). Learning from Non-Assessed Resources: Deep Multi-Type Knowledge Tracing.	Forgetting-Aware	EdNet, Junyi, MORF	AUC, RMSE, MSE	None	Visual	Not considered
64	Shen, et al. (2021). Learning Process-consistent Knowledge Tracing.	Hybrid/Meta	ASSIST12, EdNet, KT1	AUC, ACC	Q-Matrix bias	Visual	Not considered
65	Cui, et al. (2024). Leveraging Pedagogical Theories to Understand Student Learning Process with Graph-based Reasonable Knowledge Tracing.	Graph-Based	ASSIST09, ASSIST12, Junyi	AUC	Sparse data	Forgetting	Not considered

doi:10.20944/preprints202510.1845.v1

Table A1. *Cont.*

ID	Reference	Category	Dataset	Metrics	Data Quality	Stability	Interpretability
66	Zhu, et al. (2024). Meta-path structured graph pre-training for improving knowledge tracing in intelligent tutoring.	Graph-Based	ASSIST09, ASSIST17, EdNet	AUC, ACC	Sparse data	Visual	Not considered
67	Zhang, et al. (2024). MLC-DKT: A multi-layer context-aware deep knowledge tracing model.	Memory-Augmented	Junyi	AUC, ACC	None	Visual	Not considered
68	Cui, et al. (2024). Model-agnostic counterfactual reasoning for identifying and mitigating answer bias in knowledge tracing.	Memory-Augmented	ASSIST09, ASSIST17, EdNet	AUC, ACC	Imbalance	Visual	Not considered
69	He, et al. (2022). Modeling knowledge proficiency using multi-hierarchical capsule graph neural network.	Graph-Based	ASSIST09, ASSIST15, ASSIST17	AUC	Sparse data	Visual	Not considered
70	Xu, et al. (2024). Modeling Student Performance Using Feature Crosses Information for Knowledge Tracing.	Sequence Modeling	ASSIST09, ASSIST12, EdNet	AUC	None	Visual	Feature attribution / attention analysis
71	Lee, et al. (2024). MonaCoBERT: Monotonic Attention Based ConvBERT for Knowledge Tracing.	Hybrid/Meta	ASSIST09, ASSIST12, ASSIST17	AUC, RMSE	Sparse data	Visual	Feature attribution / attention analysis
72	Shen, et al. (2023). Monitoring Student Progress for Learning Process-Consistent Knowledge Tracing.	Graph-Based	ASSIST12, EdNet, KT1	AUC, ACC, RMSE, MSE	Q-Matrix bias	Visual	Not considered
73	Huang, et al. (2024). Pull together: Option-weighting-enhanced mixture-of-experts knowledge tracing.	Memory-Augmented	EdNet, Eedi, KT1	AUC, ACC, RMSE	Sparse data	None	Not considered
74	Yu, et al. (2024). RIGL: A Unified Reciprocal Approach for Tracing the Independent and Group Learning Processes.	Attentive	ASSIST12, MATH	AUC, ACC, RMSE, MAE	Imbalance	Visual	Not considered
75	Dai, et al. (2024). Self-paced contrastive learning for knowledge tracing.	Hybrid/Meta	ASSIST09, ASSIST15, KDD2010	AUC, ACC, F1	Imbalance	Visual	Not considered
76	Song & Luo (2023). SFBKT: A Synthetically Forgetting Behavior Method for Knowledge Tracing.	Forgetting-Aware	ASSIST09, ASSIST12, ASSIST17	AUC	Sparse data	None	Not considered

 doi:10.20944/preprints202510.1845.v1

Table A1. *Cont.*

ID	Reference	Category	Dataset	Metrics	Data Quality	Stability	Interpretability
77	Wu, et al. (2022). SGKT: Session graph-based knowledge tracing for student performance prediction.	Forgetting-Aware	ASSIST09, ASSIST15, Statics2011	AUC	Sparse data	Quantitative	Not considered
78	Ma, et al. (2022). SPAKT: A Self-Supervised Pre-TrAining Method for Knowledge Tracing.	Memory-Augmented	ASSIST09, ASSIST12, EdNet	AUC	None	Visual	Not considered
79	Zhu, et al. (2024). Stable Knowledge Tracing Using Causal Inference.	Hybrid/Meta	ASSIST09, ASSIST15, Statics2011	AUC, F1	None	Visual	Not considered
80	Danial Hooshyar (2024). Temporal learner modelling through integration of neural and symbolic architectures.	Hybrid/Meta	Other dataset	ACC, F1, Precision, Recall	Imbalance	Sensitivity	Not considered
81	Chen, et al. (2023). TGKT-Based Personalized Learning Path Recommendation with Reinforcement Learning.	Attentive	ASSIST09, ASSIST12	Other	None	Visual	Not considered
82	Duan, et al. (2024). Towards more accurate and interpretable model: Fusing multiple knowledge relations into deep knowledge tracing.	Hybrid/Meta	ASSIST (unspecified), EdNet, Junyi	AUC, ACC, F1	Data noise	Visual	Not considered
83	Wang, et al. (2023). What is wrong with deep knowledge tracing? Attention-based knowledge tracing.	Hybrid/Meta	ASSIST09, ASSIST15, Statics2011	AUC, F1	None	Visual	Not considered
84	Piech, et al. (2015). Deep Knowledge Tracing.	Sequence Modeling	ASSIST09	AUC, ACC	None	Visual	Not considered

Appendix B. Glossary of Acronyms

This glossary provides definitions for key acronyms and abbreviations used throughout this literature review, listed alphabetically for easy reference.

ACC	Accuracy
AFM	Additive Factor Model
AI	Artificial Intelligence
ASSIST	ASSISTments (various versions including ASSIST09, ASSIST12, ASSIST15, ASSIST17)
AUC	Area Under the Curve
BCKVMN	Bridging Concept-Key-Value Memory Networks
BERT	Bidirectional Encoder Representations from Transformers
BKT	Bayesian Knowledge Tracing
CAM	Class Activation Mapping
CAT-KT	Contrastive-Augmented Transformer for Knowledge Tracing
CGLKT	Contrastive Graph Learner for Knowledge Tracing
CNN	Convolutional Neural Network
DGMN	Dynamic Graph Memory Network
DKT	Deep Knowledge Tracing
DKTSR	Deep Knowledge Tracing via Self-Attention Residuals
DKVMN	Dynamic Key-Value Memory Network
DyKT	Dynamic Knowledge Tracing
EdNet	Education Neural Network dataset
EKT	Enhanced Knowledge Tracing
GAT	Graph Attention Network
GAT-GKT	Graph Attention Network-Graph Knowledge Tracing
GCN	Graph Convolutional Network
GIKT	Graph-based Interaction Knowledge Tracing
GKT	Graph Knowledge Tracing
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
HMN	Hierarchical Memory Network
IRT	Item Response Theory
JKT	Joint Knowledge Tracing
KT	Knowledge Tracing
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
LPKT	Learning Process-consistent Knowledge Tracing
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MMAT	Mixed Methods Appraisal Tool

MMAKT	Multi-Modal Attentive Knowledge Tracing
MOOCs	Massive Open Online Courses
NPA-KT	Neural Performance Assessment for Knowledge Tracing
PFA	Performance Factor Analysis
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAI	Responsible AI
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SAINT	Separated Self-Attentive Neural Knowledge Tracing
SAKT	Self-Attentive Knowledge Tracing
SERKT	Sequential Event Representation for Knowledge Tracing
SGKT	Session Graph Knowledge Tracing
SHAP	SHapley Additive exPlanations
Statics2011	Engineering statics course dataset from 2011
VAE	Variational Autoencoder

References

1. Xieling Chen, Di Zou, Haoran Xie, Gary Cheng, and Caixia Liu. Two decades of artificial intelligence in education. *Educational Technology & Society*, 25(1):28–47, 2022.
2. Danial Hooshyar and Marek J. Druzdzel. Memory-based dynamic bayesian networks for learner modeling: Towards early prediction of learners' performance in computational thinking. *Education Sciences*, 14(8):917, 2024. doi: 10.3390/educsci14080917.
3. Stéphan Vincent-Lancrin and Reyer Van der Vlies. Trustworthy artificial intelligence (AI) in education: Promises and challenges. *OECD education working papers*, 1(218):0_1–17, 2020.
4. Sarah Alwarthan, Nida Aslam, and Irfan Ullah Khan. An explainable model for identifying at-risk student at higher education. *IEEE Access*, 10:107649–107668, 2022.
5. Danial Hooshyar, Yueh-Min Huang, Yeongwook Yang, et al. A three-layered student learning model for prediction of failure risk in online learning. *Hum.-Centric Comput. Inf. Sci*, 12:28, 2022.
6. Miao Dai, Jui-Long Hung, Xu Du, Hengtao Tang, and Hao Li. Knowledge tracing: A review of available technologies. *Journal of Educational Technology Development and Exchange*, 14(2):1–20, 2021. doi: 10.18785/jetde.1402.01.
7. Abir Abyaa, Mohammed Khalidi Idrissi, and Samir Bennani. Learner modelling: systematic review of the literature from the last 5 years. *Educational Technology Research and Development*, 67(5):1105–1143, 2019.
8. Douglas B. Lenat, Mayank Prakash, and Mary Shepherd. Cyc: toward programs with common sense. *Communications of the ACM*, 28(8):739–752, 1985.
9. Yu Lin, Hong Chen, Wei Xia, Fan Lin, Pengcheng Wu, Zongyue Wang, and Yong Li. A comprehensive survey on deep learning techniques in educational data mining. *arXiv preprint arXiv:2309.04761*, 2023. Preprint.
10. Eleni Ilkou and Maria Koutraki. Symbolic vs sub-symbolic AI methods: Friends or enemies? In *CIKM (Workshops)*, volume 2699, 2020.
11. Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 505–513, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/bac9162b47c56fc8a4d2a519803d51b3-Abstract.html.
12. Mohammad Khajah, Robert V Lindsey, and Michael C Mozer. How deep is knowledge tracing? In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 94–101, 2016.
13. Xiaolu Xiong, Siyuan Zhao, Eric G. Van Inwegen, and Joseph E. Beck. Going deeper with deep knowledge tracing. In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 545–550, 2016.

14. Danial Hooshyar, Gustav Šír, Yeongwook Yang, Eve Kikas, Raija Hämäläinen, Tommi Kärkkäinen, Dragan Gašević, and Roger Azevedo. Towards responsible AI for education: Hybrid human-AI to confront the elephant in the room, 2025. URL https://arxiv.org/abs/2504.16148.
15. Danial Hooshyar, Roger Azevedo, and Yeongwook Yang. Augmenting deep neural networks with symbolic educational knowledge: Towards trustworthy and interpretable AI for education. *Machine Learning and Knowledge Extraction*, 6(1):593–618, 2024. doi: 10.3390/make6010029.
16. Danial Hooshyar and Yeongwook Yang. Problems with SHAP and LIME in interpretable AI for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access*, 2024.
17. Yue Gong, Joseph Beck, and Neil Heffernan. How to construct more accurate student models: Comparing and optimizing knowledge tracing and performance factor analysis. *International Journal of Artificial Intelligence in Education*, 21(1-2):27–46, 2011.
18. Andrea Aler Tubella, Marçal Mora Cantallops, and Juan Carlos Nieves. How to teach responsible AI in higher education: challenges and opportunities. *Ethics and Information Technology*, 25(3):31, 2023.
19. Florina Mihai Leta and Diane-Paula Vancea. Ethics in education: Exploring the ethical implications of artificial intelligence implementation. *Ovidius University Annals: Economic Sciences Series*, 23(1):445–451, 2023.
20. Mirka Saarela, Sachini Gunasekara, and Ayaz Karimov. The EU AI act: Implications for ethical AI in education. In *International Conference on Design Science Research in Information Systems and Technology*, pages 36–50. Springer, 2025.
21. Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.
22. Xiangyu Song, Jianxin Li, Taotao Cai, Shuiqiao Yang, Tingting Yang, and Chengfei Liu. A survey on deep learning based knowledge tracing. *Knowledge-Based Systems*, 258:110036, 2022.
23. Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1995.
24. Brett van de Sande. Properties of the bayesian knowledge tracing model. In *Proceedings of the 6th International Conference on Educational Data Mining*, pages 184–190, 2013.
25. Radek Pelánek. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User modeling and user-adapted interaction*, 27(3):313–350, 2017.
26. Ye Mao. Deep learning vs. bayesian knowledge tracing: Student models for interventions. *Journal of educational data mining*, 10(2), 2018.
27. Joseph E. Beck and Kai-min Chang. Identifiability: A fundamental problem of student modeling. *Lecture Notes in Computer Science*, 4511:137–146, 2007.
28. Ryan Shaun Baker, Albert T Corbett, and Vincent Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. *Lecture Notes in Computer Science*, 5091:406–415, 2008.
29. Philip I. Pavlik, Hao Cen, and Kenneth R. Koedinger. Performance factors analysis - a new alternative to knowledge tracing. In *International Conference on Artificial Intelligence in Education*, pages 531–538. Springer, 2009.
30. Théophile Gervet, Kenneth Koedinger, Jeff Schneider, and Tom M. Mitchell. When is deep learning the best approach to knowledge tracing. In *Proceedings of the 13th International Conference on Educational Data Mining*, pages 31–40, 2020.
31. Christopher James Maclellan, Ran Liu, and Kenneth R. Koedinger. Accounting for slipping and other false negatives in logistic models of student learning. In *Educational Data Mining*, 2015.
32. Christopher James Maclellan. Investigating the impact of slipping parameters on additive factors model parameter estimates. In *Educational Data Mining*, 2016.
33. T Merembayev, S Amirgaliyeva, and K Kozhaly. Using item response theory in machine learning algorithms for student response data. In *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, 2021.
34. Adrienne S. Kline, T. Kline, Zahra Shakeri Hossein Abad, and Joon Lee. Novel feature selection for artificial intelligence using item response theory for mortality prediction. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2020.
35. Fanglan Ma, Changsheng Zhu, and Dukui Liu. A deeper knowledge tracking model integrating cognitive theory and learning behavior. *Journal of Intelligent & Fuzzy Systems*, 2024.
36. Sevvandi Kandanaarachchi and Kate Smith-Miles. Comprehensive algorithm portfolio evaluation using item response theory. *Journal of Machine Learning Research*, 2023.

37. Sein Minn, Jill-Jênn Vie, Koh Takeuchi, Hisashi Kashima, and Feida Zhu. Interpretable knowledge tracing: Simple and efficient student modeling with causal relations. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pages 13468–13476, 2021.
38. Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021. doi: 10.1098/rsta.2020.0209.
39. Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, Laurent Callot, and Tim Januschowski. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36, 2022. doi: 10.1145/3533382.
40. Andrea Cini, Ivan Marisca, Daniele Zambon, and Cesare Alippi. Graph deep learning for time series forecasting. *ACM Computing Surveys*, 57(12):1–34, 2025. doi: 10.1145/3700680.
41. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.
42. Wilson Chango, Juan A. Lara, Rebeca Cerezo, and Cristóbal Romero. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(4):e1458, 2022.
43. Sein Minn, Yi Yu, Michel C Desmarais, Feida Zhu, and Jill-Jênn Vie. Deep knowledge tracing and dynamic student classification for knowledge component mastery. In *Proceedings of the 11th International Conference on Educational Data Mining*, pages 142–151, 2018.
44. Chun-Kit Yeung and Dit-Yan Yeung. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pages 1–10, 2018.
45. Xianqing Wang, Zetao Zheng, Jia Zhu, and Weihao Yu. What is wrong with deep knowledge tracing? attention-based knowledge tracing. *Applied Intelligence*, 53:1–20, 2022. doi: 10.1007/s10489-022-03621-1.
46. Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. Neural-symbolic learning and reasoning: A survey and interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
47. Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. In *Proceedings of the 12th International Conference on Educational Data Mining*, 2019. Reported $\sim$4.43% average AUC improvement over SOTA models.
48. Yang Yang, Jian Shen, Yanru Qu, Yunfei Liu, Kerong Wang, Yaoming Zhu, Weinan Zhang, and Yong Yu. Gikt: a graph-based interaction model for knowledge tracing. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 299–315. Springer, 2020.
49. Charl Maree, Jan Erik Modal, and Christian W. Omlin. Towards responsible AI for financial transactions. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 16–21, 2020. doi: 10.1109/SSCI47803.2020.9308456.
50. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.
51. Ray Eitel-Porter. Beyond the promise: implementing ethical AI. *AI and Ethics*, 1(1):73–80, 2021. doi: 10.1007/s43681-020-00011-6.
52. Karl Werder, Balasubramaniam Ramesh, and Rongen Zhang. Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems (TMIS)*, 13(2):1–23, 2022. doi: 10.1145/3503488.
53. Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. How different groups prioritize ethical values for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 310–323, 2022.
54. Sabrina Goellner, Marina Tropmann-Frick, and Boštjan Brumen. Responsible artificial intelligence: A structured literature review. *arXiv preprint arXiv:2403.06910*, 2024. Preprint.
55. A. Tato and R. Nkambou. Infusing expert knowledge into a deep neural network using attention mechanism for personalized learning environments. *Frontiers in Artificial Intelligence*, 2022. doi: 10.3389/frai.2022.921476.

56. Chaoran Cui, Hebo Ma, Xiaolin Dong, Chen Zhang, Chunyun Zhang, Yumo Yao, Meng Chen, and Yuling Ma. Model-agnostic counterfactual reasoning for identifying and mitigating answer bias in knowledge tracing. *Neural Networks*, 178:106495, 2024. doi: 10.1016/j.neunet.2024.106495.
57. Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. Understanding rare spurious correlations in neural networks. *arXiv preprint arXiv:2202.05189*, 2022. Preprint.
58. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. doi: 10.1145/3457607.
59. Cristina Conati, Kaśka Porayska-Pomsta, and Manolis Mavrikis. AI in education needs interpretable machine learning: Lessons from open learner modelling. *arXiv preprint arXiv:1807.00154*, 2018. Preprint.
60. Ryan S. Baker and Aaron Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41, 2022. doi: 10.1007/s40593-021-00285-9.
61. Linqing Li and Zhifeng Wang. Calibrated q-matrix-enhanced deep knowledge tracing with relational attention mechanism. *Applied Sciences*, 13(4):2541, 2023. doi: 10.3390/app13042541.
62. Changqin Huang, Hangjie Wei, Qionghao Huang, Fan Jiang, Zhongmei Han, and Xiaodi Huang. Learning consistent representations with temporal and causal enhancement for knowledge tracing. *Expert Systems with Applications*, 239:123128, 2024. doi: 10.1016/j.eswa.2023.123128.
63. Zhongyi He, Wang Li, and Yonghong Yan. Modeling knowledge proficiency using multi-hierarchical capsule graph neural network. *Applied Intelligence*, 52(8):8847–8860, 2022. doi: 10.1007/s10489-021-02765-w.
64. Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. Explainable artificial intelligence in education. *Computers and Education*, 153:103912, 2022.
65. Mirka Saarela, Ville Heilala, Päivikki Jääskelä, Anne Rantakaulio, and Tommi Kärkkäinen. Explainable student agency analytics. *IEEE Access*, 9:137444–137459, 2021.
66. Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.
67. Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
68. Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.
69. Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.
70. Joakim Linja, Joonas Hämäläinen, Paavo Nieminen, and Tommi Kärkkäinen. Feature selection for distance-based regression: An umbrella review and a one-shot wrapper. *Neurocomputing*, 518:344–359, 2023. doi: 10.1016/j.neucom.2022.10.024.
71. Shuanghong Shen, Qi Liu, Zhenya Huang, Yu Zheng, Mingyu Yin, Minjuan Wang, and Enhong Chen. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, 17:1858–1879, 2024. doi: 10.1109/TLT.2024.3360169.
72. Yujing Bai, Jiaqi Zhao, Tian Wei, Qiang Cai, and Lei He. A survey of explainable knowledge tracing. *arXiv preprint arXiv:2403.07279*, 2024. URL https://arxiv.org/abs/2403.07279. Preprint.
73. Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal*, 372:n71, 2021. doi: 10.1136/bmj.n71.
74. Neal R Haddaway, Alexandra M Collins, Deborah Coughlin, Stuart Kirk, and K. B. Wray. The role of google scholar in evidence reviews and its applicability to grey literature searching. *PLoS ONE*, 10(9):e0138237, 2015. doi: 10.1371/journal.pone.0138237.
75. Quan Nha Hong, Sergi Fàbregues, Gillian Bartlett, Felicity Boardman, Margaret Cargo, Pierre Dagenais, Marie-Pierre Gagnon, Frances Griffiths, Belinda Nicolau, Alicia O'Cathain, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, 34(4):285–291, 2018. doi: 10.3233/EFI-180221.
76. Fred Paas, Alexander Renkl, and John Sweller. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1):1–4, 2003. doi: 10.1207/S15326985EP3801_1.
77. Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010. doi: 10.1109/TSMCC.2010.2053532.

78. Hengyuan Zhang, Zitao Liu, Chenming Shang, Dawei Li, and Yong Jiang. A question-centric multi-experts contrastive learning framework for improving the accuracy and interpretability of deep sequential knowledge tracing models. *ACM Transactions on Knowledge Discovery from Data*, 2024. doi: 10.1145/3674840.
79. Sribala Vidyadhari Chinta, Zichong Wang, Zhipeng Yin, N Hoang, Matthew Gonzalez, Tai Le Quy, and Wenbin Zhang. Fairaied: Navigating fairness, bias, and ethics in educational AI applications. *arXiv preprint arXiv:2407.18745*, 2024. Preprint.
80. Zhiyu Chen, Wei Ji, Jing Xiao, and Zitao Liu. Personalized knowledge tracing through student representation reconstruction and class imbalance mitigation. *arXiv preprint arXiv:2409.06745*, 2024. doi: 10.48550/arXiv.2409.06745. Preprint.
81. Tarid Wongvorachan, Okan Bulut, Joyce Xinle Liu, and Elisabetta Mazzullo. A comparison of bias mitigation techniques for educational classification tasks using supervised machine learning. *Information*, 15(6):326, 2024.
82. John Saint, Dragan Gašević, Wannisa Matcha, Nora'ayu Ahmad Uzir, and Alejandro Pardo. Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge*, pages 107–116, 2020. doi: 10.1145/3375462.3375487.
83. Seif Gad, Sherif M. Abdelfattah, and Ghodai M. Abdelrahman. Temporal graph memory networks for knowledge tracing. *arXiv preprint arXiv:2410.01836*, 2024. Preprint.
84. Younyoung Choi and Robert J. Mislevy. Evidence centered design framework and dynamic bayesian network for modeling learning progression in online assessment system. *Frontiers in Psychology*, 13:742956, 2022. doi: 10.3389/fpsyg.2022.742956.
85. Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798, 2012.
86. Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.
87. Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018.
88. Zitao Liu, Teng Guo, Qianru Liang, Mingliang Hou, Bojun Zhan, Jiliang Tang, Weiqi Luo, and Jian Weng. pykt: A python library to benchmark deep learning based knowledge tracing models. *arXiv preprint arXiv:2206.11460*, 2022. URL https://arxiv.org/abs/2206.11460. Preprint.
89. James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
90. Nancy A. Obuchowski. Nonparametric analysis of clustered ROC curve data. *Biometrics*, 53(2):567–578, June 1997. ISSN 0006-341X. doi: 10.2307/2533958.
91. Qing Li, Xin Yuan, Sannyuya Liu, Lu Gao, Tianyu Wei, Xiaoxuan Shen, and Jianwen Sun. A genetic causal explainer for deep knowledge tracing. *IEEE Transactions on Evolutionary Computation*, 2024. doi: 10.1109/TEVC.2023.3286666.
92. Yan Cheng, Gang Wu, Haifeng Zou, Pin Luo, and Zhuang Cai. A knowledge query network model based on rasch model embedding for personalized online learning. *Frontiers in Psychology*, 2022. doi: 10.3389/fpsyg.2022.846621.
93. Weizhong Zhao, Jun Xia, Xingpeng Jiang, and Tingting He. A novel framework for deep knowledge tracing via gating-controlled forgetting and learning mechanisms. *Information Processing & Management*, 2023. doi: 10.1016/j.ipm.2022.103114.
94. Hengyu Liu, Tiancheng Zhang, Fan Li, Minghe Yu, and Ge Yu. A probabilistic generative model for tracking multi-knowledge concept mastery probability. *Frontiers of Computer Science*, 2024. doi: 10.1007/s11704-023-3008-x.
95. Sannyuya Liu, Jianwei Yu, Qing Li, Ruxia Liang, Yunhan Zhang, Xiaoxuan Shen, and Jianwen Sun. Ability boosted knowledge tracing. *Information Sciences*, 2022. doi: 10.1016/j.ins.2022.02.044.
96. Song Cheng, Qi Liu, Enhong Chen, Kai Zhang, Zhenya Huang, Yu Yin, Xiaoqing Huang, and Yu Su. Adaptkt: A domain adaptable method for knowledge tracing. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022. doi: 10.1145/3488560.3498379.

97. Deliang Wang, Yu Lu, Zhi Zhang, and Penghe Chen. An efficient and generic method for interpreting deep learning based knowledge tracing models. *IEEE International Conference on Consumer Electronics*, 2023. doi: 10.58459/icce.2023.945.
98. Huazheng Luo, Zhichang Zhang, Lingyun Cui, Ziqin Zhang, and Yali Liang. An efficient state-aware coarse-fine-grained model for knowledge tracing. *Knowledge-Based Systems*, 2024. doi: 10.1016/j.knosys.2024.112375.
99. Shuanghong Shen, Zhenya Huang, Qi Liu, Yu Su, Shijin Wang, and Enhong Chen. Assessing student's dynamic knowledge state by exploring the question difficulty effect. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022. doi: 10.1145/3477495.3531939.
100. Haoxin Xu, Jiaqi Yin, Changyong Qi, Xiaoqing Gu, Bo Jiang, and Longwei Zheng. Bridging the vocabulary gap: Using side information for deep knowledge tracing. *Applied Sciences*, 2024. doi: 10.3390/app14198927.
101. Shuaishuai Zu, Li Li, and Jun Shen. CAKT: Coupling contrastive learning with attention networks for interpretable knowledge tracing. In *IEEE International Joint Conference on Neural Network*, 2023. doi: 10.1109/IJCNN54540.2023.10191799.
102. Aritra Ghosh, N. Heffernan, and Andrew S. Lan. Context-aware attentive knowledge tracing. In *Knowledge Discovery and Data Mining*, 2020. doi: 10.1145/3394486.3403282.
103. Ghodai Abdelrahman, Qing Wang, AI Shaafi, S Ahmed, FT Sithil, and IEEE. Deep graph memory networks for forgetting-robust knowledge tracing. *IEEE Transactions on Learning Technologies*, 2023. doi: 10.1109/TLT.2023.3260220.
104. Emiko Tsutsumi, Yiming Guo, Ryo Kinoshita, Maomi Ueno, S Yang, and L Guo. Deep knowledge tracing incorporating a hypernetwork with independent student and item networks. *IEEE Transactions on Learning Technologies*, 2024. doi: 10.1109/TLT.2024.3369034.
105. Shanghui Yang, Xin Liu, Hang Su, Mengxia Zhu, and Xuesong Lu. Deep knowledge tracing with learning curves. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022. doi: 10.1109/ICDMW58026.2022.00046.
106. Dongqi Wang, Liping Zhang, Yubo Zhao, Yawen Zhang, Sheng Yan, and Min Hou. Deep knowledge tracking integrating programming exercise difficulty and forgetting factors. *International Conference on Intelligent Computing*, 2024. doi: 10.1007/978-981-97-5678-0_17.
107. Emiko Tsutsumi, Tetsurou Nishio, and Maomi Ueno. Deep-irt with a temporal convolutional network for reflecting students' long-term history of ability data. In *International Conference on Artificial Intelligence in Education*, 2024. doi: 10.1007/978-3-031-64302-6_18.
108. Yu Lu, Deliang Wang, Penghe Chen, and Zhi Zhang. Design and evaluation of trustworthy knowledge tracing model for intelligent tutoring system. *IEEE Transactions on Learning Technologies*, 2024. doi: 10.1109/TLT.2024.3403135.
109. Hongyun Wang, Leilei Shi, Zixuan Han, Lu Liu, Xiang Sun, and Furqan Aziz. DF-EGM: Personalised knowledge tracing with dynamic forgetting and enhanced gain mechanisms. In *International Conference on Networking, Sensing and Control*, 2024. doi: 10.1109/ICNSC62968.2024.10760077.
110. Seounghun Kim, Woojin Kim, HeeSeok Jung, and Hyeoncheol Kim. Dikt: Dichotomous knowledge tracing. In *International Conference on Intelligent Tutoring Systems*, 2021. doi: 10.1007/978-3-030-80421-3_5.
111. Linhao Zhou, Sheng hua Zhong, and Zhijiao Xiao. Discovering multi-relational integration for knowledge tracing with retentive networks. *International Conference on Multimedia Retrieval*, 2024. doi: 10.1145/3652583.3658030.
112. Wei Zhang, Zhongwei Gong, Peihua Luo, and Zhixin Li. DKVMN-KAPS: Dynamic key-value memory networks knowledge tracing with students' knowledge-absorption ability and problem-solving ability. *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.3388718.
113. Fei Wang, Zhenya Huang, Qi Liu, Enhong Chen, Yu Yin, Jianhui Ma, and Shijin Wang. Dynamic cognitive diagnosis: An educational priors-enhanced deep knowledge tracing perspective. *IEEE Transactions on Learning Technologies*, 2023. doi: 10.1109/TLT.2023.3254544.
114. Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. EKT: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2019. doi: 10.1109/TKDE.2019.2924374.
115. Yanjun Pu, Fang Liu, Rongye Shi, Haitao Yuan, Ruibo Chen, Tianhao Peng, and Wenjun Wu. ELAKT: Enhancing locality for attentive knowledge tracing. *ACM Trans. Inf. Syst.*, 2024. doi: 10.1145/3652601.
116. S. Cai, L. Li, and LM Watterson. Enhanced knowledge tracing via frequency integration and order sensitivity. *Lecture Notes in Computer Science*, 2025. doi: 10.1007/978-981-96-0116-5_34.

117. Ming Yin and Ruihe Huang. Enhanced knowledge tracing with learnable filter. In *2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI)*, 2024. doi: 10.1109/CCAI61966.2024.10602966.
118. Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Boyu Gao, Weiqing Luo, and Jian Weng. Enhancing deep knowledge tracing with auxiliary tasks. In *The Web Conference*, 2023. doi: 10.1145/3543507.3583866.
119. Xiaopeng Guo, Zhijie Huang, Jie Gao, Mingyu Shang, Maojing Shu, and Jun Sun. Enhancing knowledge tracing via adversarial training. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2021. doi: 10.1145/3474085.3475554.
120. Yan Zhao, Huanhuan Ma, Wentao Wang, Weiwei Gao, Fanyi Yang, and Xiangchun He. Exploiting multiple question factors for knowledge tracing. *Expert systems with applications*, 2023. doi: 10.1016/j.eswa.2023.119786.
121. Juli Zhang, Ruoheng Xia, Qiguang Miao, and Quan Wang. Explore bayesian analysis in cognitive-aware key-value memory networks for knowledge tracing in online learning. *Expert systems with applications*, 2024. doi: 10.1016/j.eswa.2024.124933.
122. Jinze Wu, Zhenya Huang, Qi Liu, Defu Lian, Hao Wang, Enhong Chen, Haiping Ma, and Shijin Wang. Federated deep knowledge tracing. *Web Search and Data Mining*, 2021. doi: 10.1145/3437963.3441747.
123. Danial Hooshyar, Yueh-Min Huang, and Yeongwook Yang. Gamedkt: Deep knowledge tracing in educational games. *Expert systems with applications*, 2022. doi: 10.1016/j.eswa.2022.116670.
124. Zhijie Liang, Ruixia Wu, Zhao Liang, Juan Yang, Ling Wang, and Jianyu Su. GELT: A graph embeddings based lite-transformer for knowledge tracing. *PLoS ONE*, 2024. doi: 10.1371/journal.pone.0301714.
125. Wei Yanyou, Guan Zheng, Wang Xue, Yan Yu, and Yang Zhijun. Global feature-guided knowledge tracing. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, 2024. doi: 10.1145/3651671.3651763.
126. Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. Graph-based knowledge tracing: Modeling student proficiency using graph neural networks. In *IEEE/WIC/ACM International Conference on Web Intelligence*, 2019. doi: 10.3233/WEB-210458.
127. Xinhua Wang, Shasha Zhao, Lei Guo, Lei Zhu, Chaoran Cui, and Liancheng Xu. Graphca: Learning from graph counterfactual augmentation for knowledge tracing. *IEEE/CAA Journal of Automatica Sinica*, 2023. doi: 10.1109/JAS.2023.123678.
128. Sannyuya Liu, Shengyingjie Liu, Zongkai Yang, Jianwen Sun, Xiaoxuan Shen, Qing Li, Rui Zou, and Shangheng Du. Heterogeneous evolution network embedding with temporal extension for intelligent tutoring systems. *IEEE Transactions on Learning Technologies*, 2024. doi: 10.1109/TLT.2024.3381054.
129. Haiqin Yang and L. Cheung. Implicit heterogeneous features embedding in deep knowledge tracing. *Cognitive Computation*, 2018. doi: 10.1007/s12559-017-9522-0.
130. Jiahao Chen, Zitao Liu, Shuyan Huang, Qiongqiong Liu, and Weiqing Luo. Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. In *AAAI Conference on Artificial Intelligence*, 2023. doi: 10.48550/arXiv.2302.06885.
131. Jia Xu, Xinyu Huang, Teng Xiao, and Pin Lv. Improving knowledge tracing via a heterogeneous information network enhanced by student interactions. *Expert systems with applications*, 2023. doi: 10.1016/j.eswa.2023.120853.
132. Shilong Shu, Liting Wang, and Junhua Tian. Improving knowledge tracing via considering students' interaction patterns. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2024. doi: 10.1007/978-981-97-2259-4_30.
133. Ting Long, Jiarui Qin, Jian Shen, Weinan Zhang, Wei Xia, Ruiming Tang, Xiuqiang He, and Yong Yu. Improving knowledge tracing with collaborative information. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022. doi: 10.1145/3488560.3498374.
134. M. Zhang, X. Zhu, Y. Ji, J Li, MG Newman, and JZ Wang. Input-aware neural knowledge tracing machine. *Lecture notes in computer science*, 2021. doi: 10.1007/978-3-030-68799-1_25.
135. Liangliang He, Xiao Li, Pancheng Wang, Jintao Tang, and Ting Wang. Integrating fine-grained attention into multi-task learning for knowledge tracing. *World wide web (Bussum)*, 2023. doi: 10.1007/s11280-023-01190-y.
136. Zhanxuan Chen, Zhengyang Wu, Qiuying Ye, and Yunxuan Lin. Interaction sequence temporal convolutional based knowledge tracing. *Lecture notes in computer science*, 2024. doi: 10.1007/978-981-97-5615-5_36.
137. Jianwen Sun, Fenghua Yu, Qian Wan, Qing Li, Sannyuya Liu, and Xiaoxuan Shen. Interpretable knowledge tracing with multiscale state representation. In *The Web Conference*, 2024. doi: 10.1145/3589334.3645373.

138. Hanshuang Tong, Zhen Wang, Yun Zhou, Shiwei Tong, Wenyuan Han, and Qi Liu. Introducing problem schema with hierarchical exercise graph for knowledge tracing. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022. doi: 10.1145/3477495.3532004.
139. Xiangyu Song, Jianxin Li, Yifu Tang, Taige Zhao, Yunliang Chen, and Ziyu Guan. JKT: A joint graph convolutional network based deep knowledge tracing. *Information Sciences*, 2021. doi: 10.1016/j.ins.2021.08.100.
140. Jianguo Pan, Zhengyang Dong, Lijun Yan, and Xia Cai. Knowledge graph and personalized answer sequences for programming knowledge tracing. *Applied Sciences*, 2024. doi: 10.3390/app14177952.
141. Wenbin Gan, Yuan Sun, Yi Sun, and HG He. Knowledge interaction enhanced sequential modeling for interpretable learner knowledge diagnosis in intelligent tutoring systems. *Neurocomputing*, 2022. doi: 10.1016/j.neucom.2022.02.080.
142. Jinseok Lee and D. Yeung. Knowledge query network for knowledge tracing: How knowledge interacts with skills. *International Conference on Learning Analytics and Knowledge*, 2019. doi: 10.1145/3303772.3303786.
143. Wenbin Gan, Yuan Sun, Yi Sun, JM Jiang, and Y Liu. Knowledge structure enhanced graph representation learning model for attentive knowledge tracing. *International Journal of Intelligent Systems*, 2022. doi: 10.1002/int.22819.
144. Yongkang Xiao, Rong Xiao, Ning Huang, Yixin Hu, Huan Li, Bo Sun, and AK Roy. Knowledge tracing based on multi-feature fusion. *Computing and Informatics*, 2023.
145. Bihan Xu, Zhenya Huang, Jia-Yin Liu, Shuanghong Shen, Qi Liu, Enhong Chen, Jinze Wu, and Shijin Wang. Learning behavior-oriented knowledge tracing. In *Knowledge Discovery and Data Mining*, 2023. doi: 10.1145/3580305.3599407.
146. Chunpai Wang, Siqian Zhao, and Shaghayegh Sherry Sahebi. Learning from non-assessed resources: Deep multi-type knowledge tracing. *Educational Data Mining*, 2021.
147. Shuanghong Shen, Qi Liu, Enhong Chen, Zhenya Huang, Wei Huang, Yu Yin, Yu Su, and Shijin Wang. Learning process-consistent knowledge tracing. In *Knowledge Discovery and Data Mining*, 2021. doi: 10.1145/3447548.3467237.
148. Jiajun Cui, Hong Qian, Bo Jiang, and Wei Zhang. Leveraging pedagogical theories to understand student learning process with graph-based reasonable knowledge tracing. In *Knowledge Discovery and Data Mining*, 2024. doi: 10.1145/3637528.3671853.
149. Menglin Zhu, Liqing Qiu, and Jingcheng Zhou. Meta-path structured graph pre-training for improving knowledge tracing in intelligent tutoring. *Expert systems with applications*, 2024. doi: 10.1016/j.eswa.2024.124451.
150. Suojuan Zhang, Jie Pu, Jing Cui, Shuanghong Shen, Weiwei Chen, Kun Hu, and Enhong Chen. MLC-DKT: A multi-layer context-aware deep knowledge tracing model. *Knowledge-Based Systems*, 2024. doi: 10.1016/j.knosys.2024.112384.
151. Lixiang Xu, Zhanlong Wang, Suojuan Zhang, Xin Yuan, Minjuan Wang, and Enhong Chen. Modeling student performance using feature crosses information for knowledge tracing. *IEEE Transactions on Learning Technologies*, 2024. doi: 10.1109/TLT.2024.3381045.
152. Unggi Lee, Yonghyun Park, Yujin Kim, Seongyune Choi, and Hyeoncheol Kim. Monacobert: Monotonic attention based convbert for knowledge tracing. *Lecture notes in computer science*, 2024. doi: 10.1007/978-3-031-63031-6_10.
153. Shuanghong Shen, Enhong Chen, Qi Liu, Zhenya Huang, Wei Huang, Yu Yin, Yu Su, and Shijin Wang. Monitoring student progress for learning process-consistent knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, 2023. doi: 10.1109/TKDE.2022.3221985.
154. Tao Huang, Xinjia Ou, Huali Yang, Shengze Hu, Jing Geng, Zhuoran Xu, and Zongkai Yang. Pull together: Option-weighting-enhanced mixture-of-experts knowledge tracing. *Expert systems with applications*, 2024. doi: 10.1016/j.eswa.2024.123419.
155. Xiaoshan Yu, Chuan Qin, Dazhong Shen, Shangshang Yang, Haiping Ma, Hengshu Zhu, and Xingyi Zhang. RIGL: A unified reciprocal approach for tracing the independent and group learning processes. In *Knowledge Discovery and Data Mining*, 2024. doi: 10.1145/3637528.3671711.
156. Huan Dai, Yue Yun, Yupei Zhang, Rui An, Wenxin Zhang, and Xuequn Shang. Self-paced contrastive learning for knowledge tracing. *Neurocomputing*, 2024. doi: 10.1016/j.neucom.2024.128366.
157. Qi Song and Wenjie Luo. SFBKT: A synthetically forgetting behavior method for knowledge tracing. *Applied Sciences*, 2023. doi: 10.3390/app13137704.

158. Zhengyang Wu, Li Huang, Qionghao Huang, Changqin Huang, Yong Tang, M Ahmad, and S Alanazi. SGKT: Session graph-based knowledge tracing for student performance prediction. *Expert Systems with Applications*, 2022. doi: 10.1016/j.eswa.2022.117681.
159. Yuling Ma, Peng Han, Huiyan Qiao, Chaoran Cui, Yilong Yin, Dehu Yu, and AS García. SPAKT: A self-supervised pre-training method for knowledge tracing. *IEEE Access*, 2022.
160. Jia Zhu, Xiaodong Ma, and Changqin Huang. Stable knowledge tracing using causal inference. *IEEE Transactions on Learning Technologies*, 2024. doi: 10.1109/TLT.2023.3264772.
161. Zhanxuan Chen, Zhengyang Wu, Yong Tang, and Jinwei Zhou. TGKT-based personalized learning path recommendation with reinforcement learning. *Knowledge Science, Engineering and Management*, 2023. doi: 10.1007/978-3-031-40289-0_27.
162. Zhiyi Duan, Xiaoxiao Dong, Hengnian Gu, Xiong Wu, Zhen Li, and Dongdai Zhou. Towards more accurate and interpretable model: Fusing multiple knowledge relations into deep knowledge tracing. *Expert Systems with Applications*, 2024. doi: 10.1016/j.eswa.2023.122033.