# The Application of Similarity-Based Attention in Transformer Architectures for Knowledge Tracing: A Paradigm Shift Towards Collaborative Modeling

## 1. Introduction: The Evolution of Knowledge Tracing and the Transformer's Role

### 1.1 The Knowledge Tracing Problem: From Foundations to Limitations

Knowledge tracing (KT) is a pivotal task within educational data mining, focused on inferring a student's evolving knowledge state and accurately predicting their future performance on learning exercises.[1] For decades, the field has been dominated by two primary methodological paradigms. The first is

**Bayesian Knowledge Tracing (BKT)**, a probabilistic approach that models a student's mastery of skills using a Hidden Markov Model (HMM).[4] BKT represents the latent knowledge state as a set of binary variables, where a student either knows or does not know a single concept.[4] While celebrated for its interpretability, BKT is built on rigid assumptions, such as the initial formulation that once a skill is learned, it is never forgotten, which is a significant oversimplification of human learning.[4] This approach often struggles to capture the complex, non-linear dynamics of a student's learning process.

The advent of deep learning in educational data mining led to the development of the second major paradigm, **Deep Knowledge Tracing (DKT)**.[4] DKT, a pioneering deep learning-based KT (DLKT) model, utilizes Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, to model the student's knowledge state as a continuous hidden vector that evolves over time.[4] This approach, being purely data-driven, demonstrated substantial improvements in predictive performance and the capacity to capture complex patterns that were beyond the reach of BKT's constrained functional forms.[4] However, subsequent analyses have indicated that DKT is not universally superior to BKT and carries its own set of challenges, including issues with parallel computing, a lack of transparency, and limitations in its storage capacity due to the scalar cell state design of LSTMs.[7] The evolution from BKT to DKT marks a fundamental shift from rigid, theory-driven models to highly flexible, data-driven ones, treating the sequence of student-question-response tuples as a form of "language" to be modeled.

### 1.2 The Rise of the Transformer in KT: A New Paradigm for Sequential Modeling

The limitations of RNNs in modeling long-range dependencies, particularly their inability to process sequences in a parallel fashion, paved the way for the adoption of the Transformer architecture in knowledge tracing.[7] The Transformer, which originated in natural language processing (NLP), has been adapted to model student performance by leveraging its state-of-the-art

**self-attention mechanism**.[7] This mechanism enables the model to weigh the importance of all past interactions in a student's sequence when making a prediction for the next one, overcoming the temporal constraints of RNN-based methods.[11]

Several notable Transformer-based models have emerged in the KT landscape, including SAKT, AKT, SAINT, and SAINT+.[10] These architectures have demonstrated great effectiveness for sequential prediction tasks, primarily by using self-attention to model the "intra-student information," which is defined as the learning patterns and history within a single student's response sequence.[11] By treating a student's interaction history as a sequence of events, these models can capture rich contextual information and learn meaningful representations of the factors affecting learning.[7] However, while highly effective for dense data, these models are still constrained by the length and quality of the student's own interaction history.[11]

### 1.3 From Intra-Student to Inter-Student Modeling: A Paradigm Shift

A significant challenge in personalized learning systems is data sparsity. For new students or those with limited interaction history, the "intra-student information" is sparse and insufficient to train a reliable model.[13] This issue mirrors a classic problem in recommender systems, where a cold-start user with no history is difficult to make recommendations for. The solution, in both domains, involves a shift from a purely individualized approach to one that leverages

**"inter-student information"**—the collective intelligence of a peer group.[5]

The central thesis of this report is that the next major conceptual leap in knowledge tracing is the integration of collaborative information. By identifying and leveraging the learning behaviors of "students who have similar question-answering experiences," a model can inform predictions for a given student, even when their own history is limited.[14] This paradigm shift addresses the fundamental limitation of data sparsity by allowing the model to draw on a richer, more extensive set of data from similar peers, providing a powerful supplement to a student's own historical sequence.

# 2. Core Concept Analysis: Collaborative Information and Similarity-Based Attention

## 2.1 Defining "Collaborative Information" in Knowledge Tracing

Within the context of knowledge tracing, "collaborative information" refers to the insights and signals derived from the learning behaviors of a group of learners, particularly those identified as similar to a target student.[5] This goes beyond the traditional intra-student focus by explicitly modeling the relationships and collective patterns that exist across a student population. This approach is motivated by the observation that learners sharing similar cognitive states often display comparable problem-solving performances.[5]

Collaborative signals can manifest in several forms, each with its own architectural implications. The most direct form involves retrieving the full question-answering sequences of peers who have a history of similar interactions.[14] A more abstract approach leverages pre-calculated or learned patterns, such as "Follow-up Performance Trends" (FPTs), that represent common learning trajectories derived from the entire student corpus.[15] These trends, while not tied to a specific individual, still represent an aggregate form of collaborative information. The efficacy of a collaborative model is therefore fundamentally dependent on the definition of what constitutes "similarity" and how these external signals are integrated.

## 2.2 The Mechanisms of Similarity-Based Attention

Traditional Transformer-based models like SAINT employ a self-attention mechanism that computes attention weights based on the relationships between tokens within a single sequence, such as a student's past interactions with exercises.[10] For collaborative knowledge tracing, this mechanism must be redefined to calculate attention based on the similarity between different students or between a student and a pre-defined learning pattern.

The core of this "similarity-based attention" involves a creative adaptation of the standard attention architecture. In a cross-attention setup, a "query" vector representing the current student's learning state can be used to query a set of "key" vectors derived from the representations of similar peers or collaborative patterns. The resulting attention score becomes a measure of semantic or behavioral similarity, which allows the model to assign higher weights to the most relevant peer interactions or patterns. The model's hidden representation for the current time step is then a weighted sum of the "value" vectors from these similar peers.[7] This process allows the model to selectively and dynamically leverage the most pertinent collaborative information, thereby enhancing its ability to make accurate predictions, particularly when the intra-student data is sparse.[13] The choice of what constitutes "similarity"—be it a simple metric on question-answering history or a complex, learned embedding—is a crucial design decision that fundamentally determines the model's capability and its computational complexity.

# 3. In-Depth Examination of Relevant Models

Several models exemplify the shift towards collaborative and similarity-based attention mechanisms. They each address the problem from a distinct architectural perspective, highlighting a growing consensus that collaborative information is a vital component for robust knowledge tracing.

## 3.1 CoKT: A Retrieval-Based Collaborative Model

The "Improving Knowledge Tracing with Collaborative Information" paper presents a novel model named **CoKT**, which is a leading example of a retrieval-based collaborative approach.[5]

The model's primary contribution is its explicit integration of both intra-student information and inter-student information simultaneously.[14] The analysis indicates that CoKT is a pioneer in this area, formally defining the problem of leveraging peer data to enhance knowledge tracing performance.

The model operates through a two-step mechanism that is described as "model-agnostic and easy-to-deploy".[14] First, it employs a retrieval mechanism to identify and fetch the question-answering sequences of peer students who have a "similar question-answering experiences" to the target student.[14] The similarity for this retrieval is quantified using a formula that incorporates inverse document frequency (IDF) weighting, a method common in information retrieval to measure the rarity of terms.[14] This process effectively creates a dynamic "collaborative context" for the target student. Second, the model uses an attention mechanism to weigh the importance of this retrieved inter-student information and integrate it with the student's own historical sequence.[14] This attention mechanism acts as a sophisticated filter, allowing the model to focus on the most relevant parts of the retrieved peer data. This approach is reminiscent of Retrieval-Augmented Generation (RAG) paradigms in large language models, where a static knowledge base is dynamically enriched with relevant context to improve output quality, demonstrating a broader convergence of architectural principles in deep learning.

## 3.2 FINER: A Pattern-Based, Similarity-Aware Attention Mechanism

The Forward-Looking Knowledge Tracing (**FINER**) model presents an alternative, pattern-based approach to collaborative tracing.[15] It addresses the issue of "correlation conflicts" in learning data by incorporating pre-identified "Follow-up Performance Trends" (FPTs).[15] An FPT represents a generalized learning pattern, such as a common sequence of correct and incorrect answers, derived from a large historical dataset.

The core of the FINER model is a "novel similarity-aware attention mechanism" that aggregates these FPTs.[15] This attention mechanism computes weights based on both the "frequency and contextual similarity" of the FPTs to the student's current learning sequence.[15] This is a powerful form of collaborative modeling because it leverages the collective behavior of the student population without needing to explicitly retrieve or reference individual peers. Instead, the model learns to identify and prioritize common and relevant learning trends, fusing this aggregated collaborative information with the student's own recent history to make more accurate predictions.[15] This approach is an effective example of a system that leverages the interactions of "similar users" by abstracting those interactions into generalizable patterns.

### 3.3 Coral: A Graph-Based Collaborative Model

The Coral model, a **Collaborative** cognitive diagnosis model, represents a third distinct architectural approach.[5] While not Transformer-based, it is a perfect example of a system that directly models student-to-student relationships to achieve a similar goal. Coral addresses the challenge of identifying implicit collaborative connections by dynamically constructing a "collaborative graph of learners".[5] The model iteratively searches for "optimal neighbors" in a context-aware manner, effectively building a network of similar students.

Once the graph is constructed, collaborative information is extracted through a node representation learning process, which is a hallmark of Graph Neural Networks (GNNs).[5] GNNs have a demonstrated ability to model complex relationships within a graph structure, and their application to knowledge tracing is a growing trend.[2] This model frames the problem not as a temporal sequence to be processed by a Transformer, but as a relational network where a student's knowledge state is influenced by their peers. This approach is valuable for showcasing a primary alternative to Transformer-based methods and for its potential to capture more nuanced relationships between learners.

## 4. Differentiating Related Approaches and Their Limitations

A thorough analysis of Transformer-based knowledge tracing requires a careful distinction between models that directly address the core problem of collaborative filtering and those that use attention mechanisms for other, albeit related, purposes. The term "attention mechanism" is a broad umbrella term, and its specific application determines a model's true relevance to the query.

### 4.1 Attention Mechanisms for Other Relations

The **RKT (Relation-Aware Self-Attention for Knowledge Tracing)** model is a potential source of confusion due to its name. While it does use a "relation-aware" attention mechanism, its focus is not on modeling relationships between students.[17] Instead, RKT’s attention models two specific types of relationships within a single student's sequence: 1) the semantic relations between exercises, which are derived from their textual content, and 2) the student's forgetting behavior, which is modeled using an exponentially decaying kernel function.[17] The attention weights in RKT serve to capture the impact of these exercise-to-exercise and time-to-time relationships, but they do not leverage a collaborative peer group.

Similarly, early Transformer-based models like **SAINT** and **AKT** employ standard self-attention mechanisms that are strictly focused on intra-student sequence modeling.[10] Their goal is to capture dependencies and contextual information within an individual's own learning history, which, as previously noted, provides an incomplete picture when the student's data is sparse.[13]

### 4.2 Models for Knowledge State Disentanglement

Another class of advanced models uses attention and other sophisticated architectures to solve different problems entirely. The **DisenKT (Hyperbolic Hypergraph Transformer with Knowledge State Disentanglement)** model uses a "hyperbolic hypergraph transformer".[18] Its primary objective is to address the issue of "entangled knowledge state embeddings" and "representation distortion" that occurs when modeling complex relationships in Euclidean space.[13] By projecting a student's response sequence into hyperbolic space, the model can learn more accurate and interpretable representations of their knowledge state.[18] While it does use "message passing between questions and students," its core contribution is the disentanglement of knowledge states via a contrastive clustering auxiliary task, not collaborative filtering.[18]

In a similar vein, the **DisKT (Disentangled Knowledge Tracing)** model introduces a "contradiction attention mechanism" to alleviate "cognitive bias" from data.[20] This mechanism is designed to handle contradictory psychology, such as a student guessing a correct answer or making a mistake on an easy question.[20] It separates a student's familiar and unfamiliar abilities and uses attention to filter out misleading information that can lead to inaccurate predictions.[20] The attention here serves as a data-cleaning and representation-refining tool for a single student's data, not a mechanism for leveraging a peer group. The careful distinction between these models is essential, as a superficial understanding of their names could lead to misapplication.

## 5. Synthesis and Comparative Analysis: Insights and

# Challenges

The analysis reveals that the problem of collaborative knowledge tracing is being addressed through several distinct architectural paradigms, each with unique strengths and limitations. These approaches demonstrate that the field is moving beyond purely intra-student sequential modeling to embrace a more holistic, population-level view of learning.

## 5.1 Comparative Framework

The following table synthesizes the key models discussed, providing a direct comparison based on their core mechanisms and the type of relationships they model.

| Model | Core Mechanism | Attention Type | Primary Relationship Modeled | Focus | Key Contribution |
|---|---|---|---|---|---|
| **CoKT** | Retrieval-Based | Cross-Attention | Student-to-Student | **Inter-Student** | First to explicitly combine intra- and inter-student information [14] |
| **FINER** | Pattern-Based | Similarity-Aware Attention | Student-to-Patterns | **Inter-Student** | Integrates follow-up performance trends (FPTs) for improved predictions [15] |
| **Coral** | Graph-Based | Node Representa | Student-to-Student | **Inter-Student** | Constructs a dynamic |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  |  | tion Learning (GNN) |  |  | collaborative graph to find optimal neighbors [5] |
| **RKT** | Transformer-Based | Self-Attention (Relation-Aware) | Exercise-to-Exercise; Temporal | Intra-Student | Incorporates exercise and forgetting relations within a single sequence [17] |
| **SAINT** | Transformer-Based | Self-Attention | Past-to-Past Interactions | Intra-Student | Pioneer of self-attention for intra-student sequence modeling [10] |
| **DisenKT** | Hyperbolic Transformer | Message Passing | Question-to-Student; Hierarchical | Intra-Student | Addresses representational distortion in hyperbolic space [18] |

## 5.2 Challenges and Open Questions

Despite the promising advancements, the collaborative KT paradigm faces several significant challenges.

- **Defining and Measuring Similarity:** The very foundation of these models rests on a clear definition of what constitutes "similarity" between students. Is a simple metric on question-answering history, as used in CoKT, sufficient, or is a more complex, learned representation, as in Coral, necessary? A flawed similarity metric could lead to the integration of irrelevant or misleading peer data, resulting in negative transfer and degraded performance.

- **Computational Scalability:** The snippets indicate that some advanced models have high computational complexity, especially on large-scale datasets.[21] Finding and processing relevant data from a peer group introduces significant overhead that can be prohibitive for real-world applications with millions of students. A key challenge lies in developing scalable architectures that can handle this without compromising predictive performance.
- **Interpretability and Trust:** A critical question for any AI model in an educational context is whether its predictions can be trusted and interpreted. While some papers mention interpretability as a goal [3], it is unclear whether the attention weights in these collaborative models can provide a clear explanation for why a particular peer's data was influential in a prediction. For a teacher or student to accept and act on a model's recommendation, they must understand the reasoning behind it.

### 5.3 Future Directions

Future research should focus on a hybrid approach that synthesizes the strengths of the models analyzed. This could involve developing models that:

1. **Use Semantic Similarity:** Integrate advanced language models or graph embeddings to derive richer, more semantic representations for student similarity. For instance, using textual content from questions [17] or code [24] could provide a more nuanced similarity metric than simple question IDs.
2. **Ensure Scalable Architectures:** Design more efficient, scalable attention mechanisms that can handle vast numbers of potential peers. This may involve exploring sparse attention or other optimization techniques.
3. **Enhance Explainability:** Create attention mechanisms that are more transparent, allowing for the visualization and interpretation of which peer interactions or collaborative patterns are deemed most relevant to a given prediction. This would not only improve trust but also provide valuable pedagogical insights.

## 6. Conclusion and Recommendations

The analysis confirms that the user's query is at the forefront of contemporary research in knowledge tracing. While no single "off-the-shelf" solution perfectly encapsulates all criteria, a new collaborative paradigm is emerging. Models like CoKT and FINER represent the most direct and promising avenues of research, moving beyond the limitations of purely intra-student modeling to leverage the power of a peer group. The success of these models, alongside the graph-based Coral, demonstrates that collaborative filtering is a powerful and necessary addition to the knowledge tracing toolkit, particularly for addressing the pervasive issue of data sparsity.

For a technical team seeking to build or research a model of this type, the following actionable recommendations are provided:

- **Architectural Study:** Focus on the architectural patterns of CoKT and FINER. CoKT's two-stage, retrieval-based approach offers a practical and modular design that could be adapted to various base models. Conversely, FINER's end-to-end, pattern-based approach is a sophisticated alternative that implicitly learns collaborative signals.
- **Data Strategy:** Prioritize access to large-scale, public datasets with a significant number of students and interactions, such as the RIIID dataset.[10] These datasets are crucial for benchmarking new models and for training the robust collaborative embeddings required.
- **Research Focus:** The core challenge lies not just in implementing a collaborative model, but in a deeper understanding of what constitutes "similarity" in a learning context. The team should allocate research resources to this fundamental problem, exploring how to define and learn meaningful student similarity.
- **Proactive Challenge Mitigation:** Be prepared to address the inherent challenges of computational complexity and interpretability. Initial prototypes may be slow or difficult to explain, and proactive efforts to design efficient algorithms and transparent visualization tools will be essential for real-world deployment.

## Works cited

1. Dtransformer: Stable Knowledge Tracing with Diagnostic Transformer - pyKT, accessed on August 22, 2025, https://pykt.org/dtransformer
2. QSDKT: Graph-Based Dynamic Knowledge Tracing through Question-Skill Similarity, accessed on August 22, 2025, https://www.researchgate.net/publication/392792626_QSDKT_Graph-Based_Dynamic_Knowledge_Tracing_through_Question-Skill_Similarity
3. Advanced Knowledge Tracing: Incorporating Process Data and Curricula Information via an Attention-Based Framework for Accuracy and Interpretability, accessed on August 22, 2025, https://jedm.educationaldatamining.org/index.php/JEDM/article/download/689/215
4. Deep Knowledge Tracing - Stanford University, accessed on August 22, 2025, https://stanford.edu/~cpiech/bio/papers/deepKnowledgeTracing.pdf
5. Improving Knowledge Tracing with Collaborative Information | Request PDF - ResearchGate, accessed on August 22, 2025, https://www.researchgate.net/publication/358627455_Improving_Knowledge_Tracing_with_Collaborative_Information
6. Deep Learning Based Knowledge Tracing: A Review, a Tool and Empirical Studies,

accessed on August 22, 2025, https://www.computer.org/csdl/journal/tk/2025/08/10933562/25d58Psv8Fa

7. Deep Knowledge Tracing with Transformers - PMC, accessed on August 22, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7334675/
8. Graph-based knowledge tracing: Modeling student proficiency using graph neural networks | Request PDF - ResearchGate, accessed on August 22, 2025, https://www.researchgate.net/publication/355766460_Graph-based_knowledge_tracing_Modeling_student_proficiency_using_graph_neural_networks
9. DKT2: Revisiting Applicable and Comprehensive Knowledge Tracing in Large-Scale Data - arXiv, accessed on August 22, 2025, https://arxiv.org/pdf/2501.14256
10. Multi-granulariy Time-based Transformer for Knowledge Tracing - arXiv, accessed on August 22, 2025, https://arxiv.org/pdf/2304.05257
11. A Multi-Layer Attention Knowledge Tracking Method with Self-Supervised Noise Tolerance, accessed on August 22, 2025, https://www.mdpi.com/2076-3417/15/15/8717
12. Towards an Appropriate Query, Key, and Value Computation for Knowledge Tracing - arXiv, accessed on August 22, 2025, https://arxiv.org/abs/2002.07033
13. Hyperbolic Hypergraph Transformer With Knowledge State ..., accessed on August 22, 2025, https://www.computer.org/csdl/journal/tk/2025/08/11003808/26JJgtINgje
14. Improving Knowledge Tracing with Collaborative ... - Weinan Zhang, accessed on August 22, 2025, https://wnzhang.net/papers/2022-wsdm-cokt.pdf
15. Advancing Knowledge Tracing by Exploring Follow-up Performance Trends - arXiv, accessed on August 22, 2025, https://arxiv.org/html/2508.08019v1
16. GRAPH-BASED KNOWLEDGE TRACING: MODELING STUDENT PROFICIENCY USING GRAPH NEURAL NET- WORK, accessed on August 22, 2025, https://rlgm.github.io/papers/70.pdf
17. RKT : Relation-Aware Self-Attention for Knowledge Tracing ..., accessed on August 22, 2025, https://www.researchgate.net/publication/343986574_RKT_Relation-Aware_Self-Attention_for_Knowledge_Tracing
18. Hyperbolic Hypergraph Transformer With Knowledge State ..., accessed on August 22, 2025, https://www.researchgate.net/publication/391747571_Hyperbolic_Hypergraph_Transformer_with_Knowledge_State_Disentanglement_for_Knowledge_Tracing
19. Yuncheng Jiang's research works | South China Normal University and other places, accessed on August 22, 2025, https://www.researchgate.net/scientific-contributions/Yuncheng-Jiang-2162960077
20. Disentangled Knowledge Tracing for Alleviating Cognitive Bias - arXiv, accessed on August 22, 2025, https://arxiv.org/abs/2503.02539
21. Evolutionary Neural Architecture Search for Transformer in Knowledge Tracing | OpenReview, accessed on August 22, 2025, https://openreview.net/forum?id=G14N38AjpU

22. Deep Knowledge Tracing Integrating Temporal Causal Inference and PINN - MDPI, accessed on August 22, 2025, https://www.mdpi.com/2076-3417/15/3/1504
23. [2502.10396] DASKT: A Dynamic Affect Simulation Method for Knowledge Tracing - arXiv, accessed on August 22, 2025, https://arxiv.org/abs/2502.10396
24. ECKT: Enhancing Code Knowledge Tracing via Large Language Models - eScholarship, accessed on August 22, 2025, https://escholarship.org/content/qt8001b5mp/qt8001b5mp_noSplash_8612d476fe311e6671d46f9dce1c5e14.pdf