KTCF: Actionable Recourse in Knowledge Tracing via Counterfactual Explanations for Education

Woojin Kim, Changkwon Lee, Hyeoncheol Kim

Department of Computer Science and Engineering, Korea University
{woojinkim1021, eckdrnjs, harrykim}@korea.ac.kr

Abstract

Using Artificial Intelligence to improve teaching and learning benefits greater adaptivity and scalability in education. Knowledge Tracing (KT) is recognized for student modeling task due to its superior performance and application potential in education. To this end, we conceptualize and investigate counterfactual explanation as the connection from XAI for KT to education. Counterfactual explanations offer actionable recourse, are inherently causal and local, and easy for educational stakeholders to understand who are often non-experts. We propose `KTCF`, a counterfactual explanation generation method for KT that accounts for knowledge concept relationships, and a post-processing scheme that converts a counterfactual explanation into a sequence of educational instructions. We experiment on a large-scale educational dataset and show our `KTCF` method achieves superior and robust performance over existing methods, with improvements ranging from 5.7% to 34% across metrics. Additionally, we provide a qualitative evaluation of our post-processing scheme, demonstrating that the resulting educational instructions help in reducing large study burden. We show that counterfactuals have the potential to advance the responsible and practical use of AI in education. Future works on XAI for KT may benefit from educationally grounded conceptualization and developing stakeholder-centered methods.

Introduction

The use of Artificial Intelligence (AI) to improve education has broad benefits for scaling personalized learning and teaching, from understanding students' learning status to automating instructional decisions (Vincent-Lancrin and Van der Vlies 2020; Nguyen et al. 2023). Under this trend, deep learning-based Knowledge Tracing (KT) has become a prominent research area, demonstrating superior performance in modeling students' knowledge mastery (Piech et al. 2015; Liu et al. 2025). KT aims to predict a student's future performance over time from previous learning history.

However, AI systems may introduce unwanted risk into education (Alfredo et al. 2024), and such concerns are recognized by policymakers. The European Union AI Act classifies educational AI models as high-risk, especially those determining access, admission, or evaluating learning outcomes (EU 2024). The Act requires AI providers to *"perform model evaluation ... with a view to identifying and mitigating systemic risks."* U.S. Department of Education further emphasizes that AI systems should *"leverage automation to advance learning outcomes while protecting human decision making and judgment"* (Office of Educational Technology 2023). In this regard, Explainable AI(XAI) is key to centering human agency for educational stakeholders and fostering trust by making predictions understandable (Khosravi et al. 2022).

Most XAI works for KT focus on model-based interpretability, either incorporating attention mechanisms or integrating educational psychometric theories (Bai et al. 2024). The former inspects the KT model's internal behaviors via attention heatmaps (Ghosh, Heffernan, and Lan 2020; Zhao et al. 2020; Qin et al. 2025), while the latter predicts psychometric parameters, such as Item Response Theory (Baker 2001), and uses those as explanations (Chen et al. 2023; Sun et al. 2024; Huang et al. 2024). While these methods address *'what'* questions, this leaves room to explore *'why?'* and *'how?'* questions (Miller 2019).

In this work, we investigate the potential of counterfactual XAI for KT and propose `KTCF`, a novel counterfactual explanation method for KT. Our counterfactual explanations are produced as actionable recourse that is causal, suggesting input changes to achieve the desired outcome (Wachter, Mittelstadt, and Russell 2017; Ustun, Spangher, and Liu 2019; Karimi, Schölkopf, and Valera 2021). We believe that counterfactual explanations are suitable for education, as they show higher user satisfaction and trust than other explanation forms (Wachter, Mittelstadt, and Russell 2017; Warren, Byrne, and Keane 2024). In KT, an example explanation is *"to change KT model's prediction on knowledge concept (KC) kc_{10} from incorrect to correct, student should change their response on previously incorrect kc_2 and kc_5."* This explanation may serve as an educational instruction that guides student actions in learning process.

Our contributions of this work are:

- We propose a counterfactual explanation generation method for KT that accounts for KC relationships and a post-processing scheme to convert a counterfactual explanation to actionable steps of educational instructions.
- We specify conceptualization, problem formulation, and desired properties of generating a counterfactual explanation for KT under the educational context.

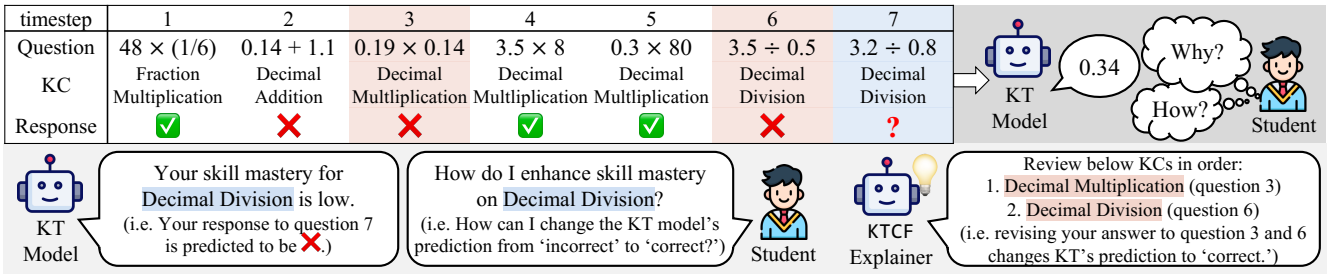


Figure 1: An application scenario of our counterfactual explanation for KT under an educational context.

An Application Scenario

To illustrate our method's contribution, we present a hypothetical counterfactual explanation in Figure 1. A student solves a series of problems on *'Decimal Arithmetic,'* and a KT model learns the student's skill mastery based on the student's learning history. The KT model predicts the probability that the student will answer correctly on *'Decimal Division'* at timestep 7 is 0.34, translated as the student's skill mastery on *'Decimal Division'* is low.

Rather than asking about what features were relevant or what sections of learning history the KT model focused on for prediction, the student would like to know how to enhance the mastery of *'Decimal Division,'* translated as a counterfactual question to achieve the desired outcome: *"how can I change the KT model's prediction to correct?"*

Our `KTCF` method identifies a sequence of incorrect KCs to be corrected in the student's previous learning history, which translates to indicating changes to the student's response that would flip the KT's prediction to the desired outcome. Thus, this explanation can guide the learning process according to the student's needs.

In Relation to Educational Theories Our approach is grounded by Bloom's Mastery Learning (Block 1971). The theory operationalizes teaching as initial instruction followed by formative assessment. Based on the diagnosis, correction or review procedures are given to failed students so that misunderstandings are not propagated. Bloom states,

> *Students respond best when diagnosis is accompanied by specific prescription of alternative instruction materials and processes they can use to overcome their learning difficulties* (Bloom 1968).

Bloom demonstrated the effects of Mastery Learning versus conventional classes (teacher-student ratio 1:30), and 1 - 1 tutoring. Results show that tutored students' achievements are 2 sigma above the conventional class average (Bloom 1984). Bloom termed the '2 Sigma Problem,' which aims to find ways to replicate the effect of 1-1 tutoring for students.

KT models may perform as a highly scalable diagnostic tool for measuring skill mastery, and our explanation method could potentially support effective correction or review procedures, resembling some aspects of 1-1 tutoring. Within this context, it is conceivable that KT and our explanation method could serve as a potential solution to the 2-sigma problem.

Related Works

Counterfactual Explanation for Education

While counterfactual explanation is widely studied in domains such as medicine, finance, and process monitoring (Wang, Samsten, and Papapetrou 2021; Wang et al. 2023a; Huang, Metzger, and Pohl 2021), its applications on education are relatively underexplored.

Counterfactual explanations for education have been applied to predicting student performance, detecting at-risk students, and analyzing dropout patterns. Afrin, Hamilton, and Thevathyan uses Diverse Counterfactual Explanations (`DiCE`) (Mothilal, Sharma, and Tan 2020) to analyze predictions on whether students will pass or fail at their course (Afrin, Hamilton, and Thevathyan 2023). Tsiakmaki and Ragos generates counterfactuals guided by the nearest class prototype for explaining predictions on whether a student will pass or fail the final exam (Tsiakmaki and Ragos 2021). Smith, Chimedza, and Bührmann generates counterfactuals by perturbing predictors generated by SHAP for explaining whether a student is at risk of failing a course (Smith, Chimedza, and Bührmann 2022). Zhang et al. generates and visualizes counterfactuals for analyzing dropout patterns in online learning (Zhang et al. 2023a).

For KT, the notion of counterfactual reasoning has been implemented as a means of aiding model prediction, but has not been discussed under the XAI context (Wang et al. 2023b; Zhang et al. 2023b; Cui et al. 2024).

Explainable Knowledge Tracing

Aforementioned, there exists two threads of XAI research for KT, model-based and post-hoc (Bai et al. 2024). Model-based explanations for KT are presented as heatmaps of attention weights (Ghosh, Heffernan, and Lan 2020; Zhao et al. 2020; Qin et al. 2025) and line plots of predicted parameters for psychometric theories (Chen et al. 2023; Sun et al. 2024; Huang et al. 2024). Post-hoc explanations for KT employ KC maps derived from KT predictions and assess local feature importance using LRP or SHAP (Lu et al. 2023; Wang et al. 2022; Valero-Leal, Carlon, and Cross 2023).

These forms of explanation allow for inspecting the internal workings of KT models and finding input-output associations. Specifically for local explanations, attention heatmaps can answer *"what part of sequence did the KT model most focus on when making a prediction?"*; psychometric line plots can answer *"does the KT model accurately*

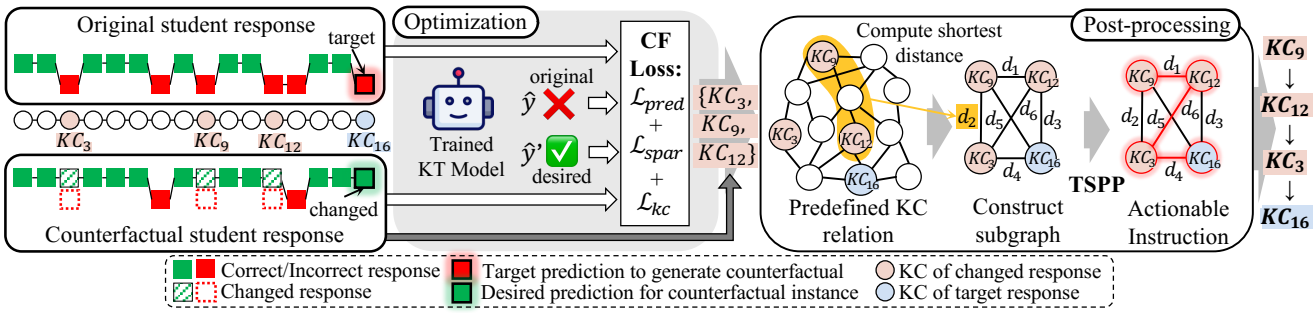


Figure 2: An illustration of our proposed KTCF method for generating counterfactual explanations for KT.

capture the student's learning process?"; and post-hoc feature relevance can answer *"what positive or negative contribution does a KC have on the KT's prediction?"*

While existing methods can provide tools for model behavior inspection and key features analysis, these do not necessarily translate to immediate and actionable educational instruction for a specific, unique student. Further, limitations of attention mechanisms for explanations have been addressed in the previous XAI literature (Serrano and Smith 2019; Bai et al. 2021; Liu et al. 2022a).

Notably, Lu et al. emphasizes stakeholders' agency in using KT for educational support, using Deep SHAP to explain KT prediction as *"kc_i is your weak skill, influenced by your performance on kc_j, kc_k"* (Lu et al. 2024). Their user study shows that explaining KT's decisions increases trust, credibility, and knowledge for students and teachers.

Counterfactual Explanations for KT

Problem Conceptualization

Guided by current advice on XAI research (Freiesleben and König 2023; Langer et al. 2021; Miller 2019), we believe that "*explanations are not just the presentation of associations and causes, they are contextual*" (Miller 2019).

We first specify what definition we follow for interpretability and formulate the goal, concepts, and purpose of counterfactual explanations in education.

We follow the definition of Murdoch et al.,

> *We define interpretable machine learning as extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model* (Murdoch et al. 2019).

Primary goal is of our explanation is to make KT model predictions understandable to educational stakeholders, which are students, teachers, families and caregivers.

For concepts, we propose *explanandum* (i.e., what is to be explained) and *explanans* (i.e., what we want to explain) for our work (Miller 2019). We define explanandum for KT as *"how the student learning history would have to be different for the KT model to predict high mastery of a specific KC?"* and explanans as *"students' previous incorrect KCs to be corrected."*

Problem Formulation

Let $X^{\text{orig}} = [(kc_1, r_1^{\text{orig}}), (kc_2, r_2^{\text{orig}}), \cdots, (kc_t, r_t^{\text{orig}})]$ be a original sequence of a student's learning history up to time T, and $R^{\text{orig}} = [r_1^{\text{orig}}, r_2^{\text{orig}}, \cdots, r_t^{\text{orig}}]$ be a sequence of student's responses of X. Let $f : X^{\text{orig}} \rightarrow [0, 1]$ be a trained KT model that predicts whether a student will give a correct answer to a KC at timestep t. We focus on a target KC at specific t, denoted as kc_{target}, where student's response is incorrect and KT model predicted the student response to be incorrect ($\hat{y}_t = f(X_t^{\text{orig}}) = 0$). kc_{target} is the last KC in this sequence. Our objective is to find a counterfactual responses $R^{\text{cf}} = [r_1^{\text{cf}}, r_2^{\text{cf}}, \cdots, r_t^{\text{cf}}]$ that constitutes a counterfactual learning history $X^{\text{cf}} = [(kc_1, r_1^{\text{cf}}), (kc_2, r_2^{\text{cf}}), \cdots, (kc_t, r_t^{\text{cf}})]$ such that KT model predicts positive ($\hat{y}_t = f(X^{\text{cf}}) = 1$) for a specific kc_{target}.

Properties of Counterfactual Explanations for KT

We propose properties that a 'good' counterfactual explanation should have under educational context:

- **Intervention Sparsity**: Counterfactual explanations should suggest minimum changes as well as be close to the original learning history (Wachter, Mittelstadt, and Russell 2017).
- **Actionability**: Counterfactual explanations should include actionable changes. It is unrealistic to suggest changing KCs to incorrect when the student already answered them correctly (Wachter, Mittelstadt, and Russell 2017; Verma et al. 2024).
- **KC Level Granularity**: Explanations should be framed in terms of KCs, not question items. A good explanation would suggest *"improve your 'Decimal Division' skill..."* rather than *"answer 'Question 17' correctly..."*
- **KC Relationship Coherence**: Explanations should consider the relationship between KCs. For instance, it would be implausible to suggest practicing *'Integer Addition'* skill to a student solving *'Algebraic Equations.'*
- **Form of Discrete, Sequential Steps of Actions**: Counterfactual explanations should be presented in a series of actions that would guide students from the current decision to the desired decision, complying with the purpose of algorithmic recourse (Verma et al. 2024).

Algorithm 1: KTCF: Counterfactual Explanation for KT

Input: A student's learning history X^{orig}, target KC kc_{target}, trained KT model f, KC relation graph G_{kc}
Parameter: $\lambda_{\text{spar}}, \lambda_{\text{kc}}$, max iterations N_{iter}, learning rate η, early stopping threshold τ
Output: An counterfactual response R^{cf} derived from the original response R^{orig}

Define mask $\mathbf{m}$ where $m_t \leftarrow \mathbb{I}_{r_t \in R^{\text{orig}}}[r_t^{\text{orig}} = 0]$ for all t
Initialize R^{cf}
for iteration from 1 to N_{iter} **do**
  $\mathcal{L}_{\texttt{KTCF}} \leftarrow \mathcal{L}_{\text{pred}} + \lambda_{\text{spar}} \cdot \mathcal{L}_{\text{spar}} + \lambda_{\text{kc}} \cdot \mathcal{L}_{\text{kc}}$
  Compute gradient: $\theta \leftarrow \nabla_{R^{\text{cf}}} \mathcal{L}_{\texttt{KTCF}}$
  Update counterfactual: $R^{\text{cf}} \leftarrow R^{\text{cf}} - \eta \cdot \theta$
  Project onto actionable elements: $R^{\text{cf}} \leftarrow R^{\text{cf}} \odot \mathbf{m} + R^{\text{orig}} \odot (1 - \mathbf{m})$
  if $\mathcal{L}_{\texttt{KTCF}} < \tau$ **then**
    break
  end if
end for
return R^{cf}

Algorithm 2: Post-processing for Sequential Actions

Input: Original response sequence R^{orig}, counterfactual response sequence R^{cf}, KC sequence $[kc_1, kc_2, \cdots, kc_t]$, KC relation graph $G_{\text{kc}} = (V_{\text{kc}}, E)$, target KC kc_{target}.
Output: Path H' derived from R^{cf}

$\mathcal{I} \leftarrow \{i \mid R^{\text{orig}} \neq R^{\text{cf}}\}$
$\mathcal{S}_{\text{kc}} \leftarrow \{kc_i \mid i \in \mathcal{I}\}$
$\overline{V}^{\text{CF}} \leftarrow \mathcal{S}_{\text{kc}} \cup \{kc_{\text{target}}\}$
Initialize distance matrix $D \in \mathbb{R}^{|\overline{V}^{\text{CF}}| \times |\overline{V}^{\text{CF}}|}$
for each $kc_i \in \overline{V}^{\text{CF}}$ **do**
  for each $kc_j \in \overline{V}^{\text{CF}}$ **do**
    $D[kc_i, kc_j] \leftarrow \textsc{Dijkstra}(G_{kc}, kc_i, kc_j)$
  end for
end for
Construct subgraph $G'_{\text{kc}} = (\overline{V}^{\text{CF}}, E', D)$, where $E' = \{(kc_i, kc_j) \mid kc_i, kc_j \in \overline{V}^{\text{CF}}, kc_i \neq kc_j\}$ and edge weights are $D[kc_i, kc_j]$
$H \leftarrow \textsc{Greedy}(G'_{\text{kc}}, kc_{\text{target}})$
$H' \leftarrow \textsc{Rev}(H)$
return H'

Methodology

Our KTCF method serves as a local, post-hoc outcome explanation, formulated as an optimization problem. The overview of our method is illustrated in Figure 2.

Counterfactual Explanation Given a student's original response $R = [r_1^{\text{orig}}, r_2^{\text{orig}}, \cdots, r_t^{\text{orig}}]$ from a learning history $X^{\text{orig}} = [(kc_1, r_1^{\text{orig}}), (kc_2, r_2^{\text{orig}}), \cdots, (kc_t, r_t^{\text{orig}})]$, we initialize our counterfactual response R^{cf}. Then, we generate our counterfactual explanation through a stochastic optimization using Adam (Kingma and Ba 2014).

Our loss function is defined as:

$$\mathcal{L}_{\texttt{KTCF}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{spar}} \cdot \mathcal{L}_{\text{spar}} + \lambda_{\text{kc}} \cdot \mathcal{L}_{\text{kc}}, \tag{1}$$

where our prediction loss $\mathcal{L}_{\text{pred}}$ is binary cross entropy between predicted probability on a counterfactual instance X^{cf} and the desired probability of 1.0,

$$\mathcal{L}_{\text{pred}} = -\log(f(X^{\text{cf}})), \tag{2}$$

and sparsity loss $\mathcal{L}_{\text{spar}}$ is Hamming distance between a original response R^{orig} and the counterfactual response R^{cf},

$$\mathcal{L}_{\text{spar}} = \sum_{t=1}^{T} \mathbb{I}(R_t^{\text{orig}} \neq R_t^{\text{cf}}). \tag{3}$$

To ensure that counterfactual explanations suggest KC changes that are pedagogically sound and closely related to the kc_{target}, we introduce a penalty term, KC loss $\mathcal{L}_{\text{kc}}$, based on path distance in the loss function. We utilize an undirected graph of predefined KC relations, $G_{\text{kc}} = (V_{\text{kc}}, E)$, where V_{kc} is a node set where each node is a KC, and E is an edge set where each edge $e \in E$ indicates a relationship exists between node KC_i and node KC_j.

The KC loss $\mathcal{L}_{\text{kc}}$ aims to penalize changes that are distant from the kc_{target} in the KC relation graph, as measured by the shortest path distance d. Formally, the $\mathcal{L}_{\text{kc}}$ is defined as,

$$\mathcal{L}_{\text{kc}} = \sum_{kc_i \in \overline{V}^{\text{CF}}} d(kc_i, kc_{\text{target}}) \tag{4}$$

where $\overline{V}^{\text{CF}}$ denotes KCs modified in the counterfactual. This guides the optimization to penalize changes to KCs less relevant to the kc_{target}, abiding by the KC relationship coherence.

We apply an actionability mask $\mathbf{m}$ as a crucial step in our process to ensure only the student's originally incorrect answers can be modified. It generates counterfactuals that are only actionable and enforces complete actionability. The full explanation generation process is described in Algorithm 1.

Post-processing for Sequential Actions After identifying a set of counterfactual KCs, we convert the explanation to a sequence of educational instructions using a variation of the Traveling Salesman Path Problem (TSPP) (Lam and Newman 2008). Our approach is described in Algorithm 2.

Given G_{kc}, the goal is to find a Hamiltonian path starting from the kc_{target} node that traverses all unique nodes in $\overline{V}^{\text{CF}}$. First, we calculate the shortest path distance between every node pair of our counterfactual KC set $\overline{V}^{\text{CF}}$ using Dijkstra's algorithm. Second, we create a new, smaller complete graph $G'_{\text{kc}} = (\overline{V}^{\text{CF}}, E', D)$ where edge weights D are the shortest path distance. Then, we find a Hamiltonian path H^* starting from the target KC in a greedy manner on G'_{kc}. The inversed path is provided as a sequence of actionable steps.

The complexity of Algorithm 2 is $O(|\overline{V^{\text{CF}}}|^2(|V_{\text{kc}}| + |E|)\log(|V_{\text{kc}}|))$. It is small because KTCF is optimized for sparsity. Under an educational context, this generates a sequential learning instruction that minimizes overall study burden of students.

Methods	Validity(↑)	Sparsity(↓)	Sparsity Rate(↓)	Actionability(↓)	Actionability Rate(↓)	Time(↓)
`Wachter-rand`	0.725±0.45	67.355±10.33	0.338±0.05	40.525±10.90	0.617±0.20	3.791±0.08
`DiCE-rand`	0.880±0.33	75.587±14.73	0.380±0.07	35.320±12.24	0.504±0.25	2.565 ±1.87
`KTCF-rn`	**0.930 ± 0.26**	**49.845 ± 8.44**	**0.250 ± 0.04**	**0.000 ± 0.00**	**0.000 ± 0.00**	3.024 ± 0.91
`KTCF-rand`	0.720 ± 0.45	53.985 ± 9.18	0.271 ± 0.05	**0.000 ± 0.00**	**0.000 ± 0.00**	4.700 ± 0.06
`KTCF-sr`	**0.930 ± 0.26**	50.075 ± 7.71	0.252 ± 0.04	**0.000 ± 0.00**	**0.000 ± 0.00**	3.065 ± 1.15
`KTCF-cc`	0.685 ± 0.47	55.635 ± 8.31	0.280 ± 0.04	**0.000 ± 0.00**	**0.000 ± 0.00**	3.772 ± 0.97
`KTCF-gs`	0.920 ± 0.27	49.920 ± 7.74	0.251 ± 0.04	**0.000 ± 0.00**	**0.000 ± 0.00**	**2.202 ± 1.52**

Table 1: Evaluation results of counterfacutal explanation generation methods `KTCF`, `Wachter`, and `DiCE` on XES3G5M test data. Best results are shown in bold; second-best are underlined.

Experiment

We show quantitative and qualitative evaluation of our `KTCF` method. Since choosing baselines and evaluation criteria for counterfactual explanations heavily depend on types of approaches and the application domain, we select landmark baselines and evaluation metrics from the previous counterfactual XAI literature on benchmarking (Mothilal, Sharma, and Tan 2020; Guidotti 2024; Moreira et al. 2025).

Baselines We compare our `KTCF` method to two baselines; `Wachter` (Wachter, Mittelstadt, and Russell 2017) and `DiCE` (Mothilal, Sharma, and Tan 2020). Although being a solid stream of counterfactual XAI, we do not consider Prototype or Nearest Unlike Neighbor approaches (Van Looveren and Klaise 2021; Delaney, Greene, and Keane 2021) that we regard the learning process of each student as idiosyncratic (Bloom 1968).

Evaluation Metrics We choose the following quantitative evaluation metrics: validity, sparsity, sparsity rate, actionability, and generation time. To foster evaluation standardization of XAI research, we mention that selected evaluation metrics cover four CO-12 properties: Continuity, Compactness, Context, and Contrastivity (Nauta et al. 2023).

Given a test dataset of N instances with each instance X_i has length T, we define evaluation metrics as follows:

1. **Validity(↑)** measures the fraction of counterfactuals returned that are actually counterfactuals,

$$\text{Validity} = \frac{\sum_{i=1}^{N} \mathbb{I}[f(X_i^{\text{cf}}) > 0.5]}{N}. \tag{5}$$

2. **Sparsity(↓)** measures the number of features that have changed,

$$\text{Sparsity} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{I}({R^{\text{orig}}}_t^{(i)} \neq {R^{\text{cf}}}_t^{(i)}). \tag{6}$$

3. **Sparsity Rate(↓)** measures the ratio of features that have changed relative to the total number of features,

$$\text{Sparsity Rate} = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{t=1}^{T} \mathbb{I}({R^{\text{orig}}}_t^{(i)} \neq {R^{\text{cf}}}_t^{(i)})}{T}. \tag{7}$$

4. **Actionability(↓)** measures unactionable changes in counterfactuals. In KT, an unactionable change occurs when a counterfactual suggests altering a correct response to incorrect,

$$\text{Actionability} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{I}[{R^{\text{orig}}}_t^{(i)} = 1 \wedge {R^{\text{cf}}}_t^{(i)} = 0]. \tag{8}$$

5. **Actionability Rate(↓)** is the proportion of all changes (Sparsity) that are unactionable,

$$\text{Actionability Rate} = \frac{\text{Actionability}}{\text{Sparsity}}. \tag{9}$$

6. **Generation Time(↓)** measures the time (in seconds) the algorithm takes to find a counterfactual.

Counterfactual Initialization Generating counterfactuals for categorical features is a known challenge in counterfactual XAI (Verma et al. 2024). The challenge is especially severe in KT, where binary student responses make counterfactual quality highly sensitive to initialization. To empirically investigate this issue, we experiment with five initialization strategies for binary sequences of student responses:

- Gaussian noise (`-rn`)

$$R^{\text{cf}} = R^{\text{orig}} + \lambda_{\text{noise}} \cdot \epsilon, \quad \epsilon_i \sim \mathcal{N}(0,1) \tag{10}$$

- Random binary (`-rand`)

$$R^{\text{cf}} = z, \quad z_i \sim \text{Bernoulli}(0.5) \tag{11}$$

- Soft relaxation (`-sr`)

$$R^{\text{cf}} = \sigma(z), \quad z_i \sim \mathcal{N}(0,1) \tag{12}$$

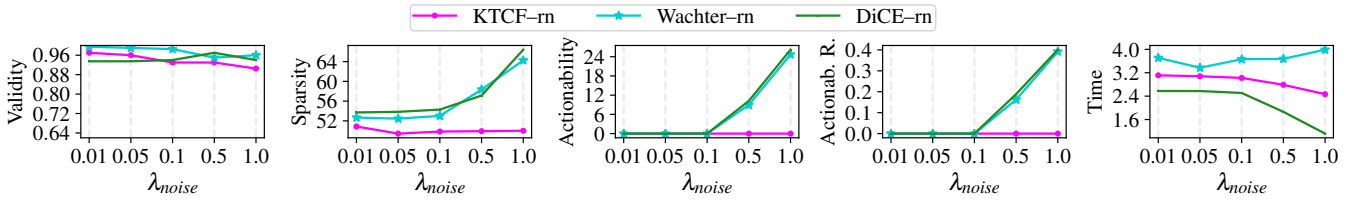
- Convex combination (`-cc`)

$$\begin{aligned} R^{\text{cf}} &= \lambda_{\text{cc}} \cdot R^{\text{orig}} + (1-\lambda_{\text{cc}}) \cdot z, \\ z_i &\sim \text{Bernoulli}(0.5), \quad \lambda_{\text{cc}} \in [0,1] \end{aligned} \tag{13}$$

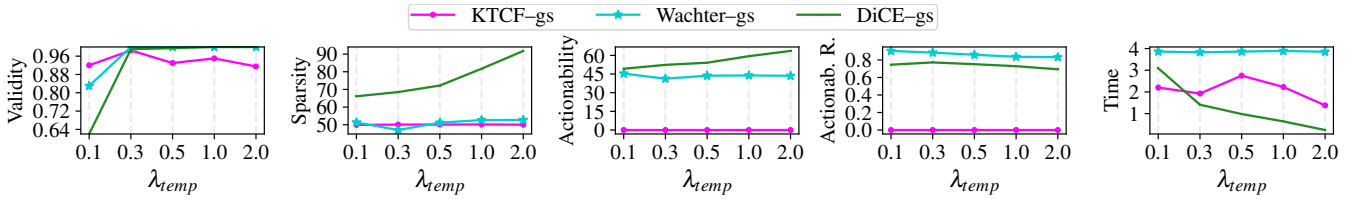
- Gumbel-Sigmoid Relaxation (Jang, Gu, and Poole 2016) (`-gs`)

$$\begin{aligned} R^{\text{cf}} &= \sigma\left(\frac{z + g_1 - g_2}{\lambda_{\text{temp}}}\right), \\ z_i \sim \mathcal{N}(0,1), &\quad g_1, g_2 \sim \text{Gumbel}(0,1), \quad \lambda_{\text{temp}} > 0 \end{aligned} \tag{14}$$

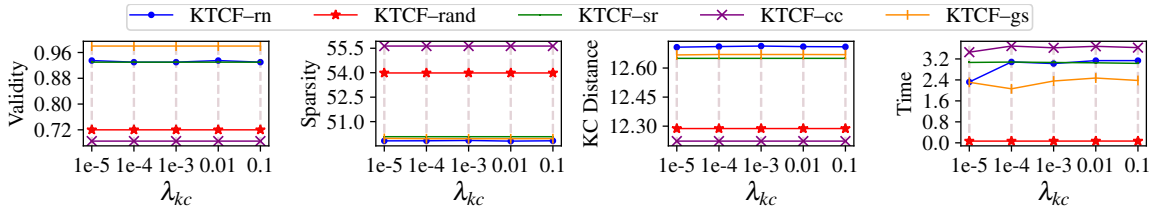
Baselines `Wachter` and `DiCE` follow random initialization(`-rand`) as indicated in their works.



(a) Comparisons of methods with Gaussian noise initialization strategy (`-rn`) with varying hyperparameter λ_{noise}.



(b) Comparisons of methods with Gumbel-Sigmoid initialization strategy (`-gs`) with varying hyperparameter λ_{temp}.



(c) Comparisons of initialization strategies on `KTCF` with varying hyperparameter λ_{kc}.

Figure 3: Ablation study on the sensitivity of `KTCF`, `Wachter`, and `DiCE` methods' performance across hyperparameter variations of initialization strategies `-rn` and `-gs`, and across hyperparameter variations of KC Loss λ_{kc}.

Dataset We use the XES3G5M (Liu et al. 2023), a large-scale educational dataset with auxiliary information on KCs. XES3G5M contains 5,549,635 interaction sequences from 18,066 students on Mathematics. The dataset also provides relationship information among KCs, which we utilize in our `KTCF` method as an undirected graph. The constructed KC relation graph has 1,175 nodes and 1,304 edges. We evaluate the approaches on 200 randomly selected instances from students with more than 45% incorrect responses. This selection was made intentionally to focus on students who could benefit from `KTCF` while minimizing bias from students with all-correct responses.

KT Architecture We use DKT (Piech et al. 2015) as our main KT architecture. DKT uses an LSTM network to model student interactions, and shows performance comparable to that of subsequent KT models featuring more advanced and complex architectures (Liu et al. 2022b). For standardizing KT research, we use the `pyKT` library (Liu et al. 2022b) for data preprocessing, training, and evaluation.

Experimental Setup and KT Performance We run all our experiments on NVIDIA GeForce RTX 4090 devices, and experiment setups are identical to `pyKT`'s setups (Liu et al. 2022b). We perform 5-fold cross-validation on DKT and report the test performance. The DKT model demonstrates 0.8415 accuracy and 0.8358 AUC for validation and 0.8253 accuracy and 0.8226 AUC for test data. For hyperparameters in Algorithm 1, we set $\lambda_{\text{spar}} = 0.1$, $\lambda_{\text{kc}} = 1\text{e}-3$, $N_{\text{iter}} = 200$, η, $\tau = 1\text{e}-4$, and $\lambda_{\text{cc}} = 0.5$.

Experiment Results

Table 1 presents our evaluation results. Overall, `KTCF` outperforms baselines across all metrics. Specifically, `KTCF-rn` improves validity by 28.3% over `Wachter-rand` and 5.7% over `DiCE-rand`, and reduces sparsity by 26.0% and 34.0%, respectively. `KTCF-gs` achieves 41.9% and 14.2% faster computation times compared to `Wachter-rand` and `DiCE-rand`. Notably, `KTCF` produces only actionable counterfactuals, fully eliminating unactionable modifications.

Our method generates valid counterfactuals with minimal, actionable changes. `KTCF-rn` and `KTCF-gs` are the most balanced in validity, sparsity, actionability, and generation time. By contrast, `Wachter` and `DiCE` fail to generate valid counterfactuals for 27.5% and 12% of students, respectively, and suggest excessive and unactionable changes.

Initialization strategies using Gaussian noise (`-rn`), soft relaxation (`-sr`), and Gumbel-Sigmoid relaxation (`-gs`) yield the best results, suggesting that soft relaxation-based initializations are most effective. In contrast, hard random (`-rand`) and convex combination (`-cc`) initializations underperform, indicating that hard random starts are ineffective for generating counterfactual explanations for KT.

Ablation Study We investigate how the choice of initialization hyperparameters influences performance, as depicted in Figure 3a and 3b. We select two initialization strategies `-rn` and `-gs` that demonstrate high performance in our final result, and compare the performance of three methods across hyperparameters λ_{noise} and λ_{temp}.

Methods	Generated Educational Instructions	Total Path Distance
Wachter	[Treemap, Ring operation cycle issues, Applications of odd and even numbers, *Integer difference times*, Find cycle rules in the calendar, **Distinguish between leap years and non-leap years**, Calendar cycles, Magic square relationship, Simple statistics table, Counting multiples, Inverted type, Prototype questions, Variation questions, Counting units of decimals]	77
Wachter *(a)*	*Integer difference times* → Calendar cycles → Ring operation cycle issues → Magic square relationship → Variation questions → Prototype questions → Inverted type → Counting multiples → Counting units of decimals → Applications of odd and even numbers → Simple statistics table → Treemap → Find cycle rules in the calendar → **Distinguish between leap years and non-leap years**	66
DiCE	[Treemap, Treemap, Ring operation cycle issues, Applications of odd and even numbers, Applications of odd and even numbers, Modulo operation, *How many changes are applied*, *The problem of two-quantity difference between dark difference type*, Find cycle rules in the calendar, Calendar cycles, Calendar cycles, Even-order magic square filling method, Magic square relationship, Simple statistics table, Counting multiples, Inverted type, Variation questions, Two-volume repulsion, Counting units of decimals, **Distinguish between leap years and non-leap years**]	95
DiCE *(a)*	*The problem of two-quantity difference between dark difference type* → Modulo operation → *How many changes are applied* → Calendar cycles → Ring operation cycle issues → Even-order magic square filling method → Magic square relationship → Counting units of decimals → Applications of odd and even numbers → Counting multiples → Variation questions → Inverted type → Two-volume repulsion → Simple statistics table → Treemap → Find cycle rules in the calendar → **Distinguish between leap years and non-leap years**	79
KTCF	[Applications of odd and even numbers, Modulo operation, **Distinguish between leap years and non-leap years**, Calendar cycles, Counting multiples]	30
KTCF *(a)*	Modulo operation → Calendar cycles → Counting multiples → Applications of odd and even numbers → **Distinguish between leap years and non-leap years**	26

Table 2: Comparisons of educational instructions generated by KTCF, Wachter, and DiCE on instance # 1,452 of the XES3G5M test data. Rows with *(a)* indicate results after applying the proposed post-processing scheme. The total path distance is the sum of the shortest paths between each consecutive pair of KCs in the explanation. Target KC k_{target} is bolded and unactionable changes are italicized. KC names are translated into English for readability; the original texts are in Chinese.

Overall, our KTCF method is highly robust across different hyperparameter values for both initialization strategies. KTCF−rn and KTCF−sg performance in validity, sparsity, and actionability remains consistently high regardless of the noise level and temperature settings. In contrast, baseline methods are highly sensitive to these hyperparameters.

To provide intuition for this result, we show the performance of KTCF on initialization strategies across ranges of hyperparameter λ_{kc} in Figure 3c. The results indicate that the KC loss itself strengthens robustness, making performance consistent to initialization strategies across all values of λ_{kc}.

Qualitative Analysis of Actionable Steps To evaluate our post-processing scheme for converting counterfactuals to sequential actions, we present an actual instance of generated educational instructions for qualitative comparison in Table 2. For KTCF, we used top performing KTCF−rn method to generate instructional steps. For DiCE, we selected the first explanation among its diverse explanations.

The total path distance is reduced after applying the post-processing scheme for all methods, indicating that our scheme effectively finds a path that reduces overall study burden of the educational instruction. KTCF provides the fewest educational instructions with complete actionable suggestions. Though average KC distance for KTCF are larger (Wachter:DiCE:KTCF=4.71:4.64:5.2), total study burden is lower(66:79:26) due to KTCF's superior sparsity.

In Table 2, KCs presented in KTCF are closely related to the target KC of *'Distinguishing between leap and non-leap years.'* To solve this problem, knowledge on divisibility rule, calendar cycles, modulo operations, and basic integer arithmetic are required, which are included in KTCF's instruction. However, instructional steps from Wachter and DiCE includes seemingly unrelated KCs to *'Distinguishing between leap and non-leap years,'* such as *'Treemap,' 'Counting units of decimals,' 'Magic square relationship,'* and *'Simple statics table.'*

Conclusion

In this work, we propose KTCF, a novel method for generating counterfactual explanations in KT as well as method for converting explanation to a sequence of steps for actionable recourse in education. Our experiments show that KTCF produces high-quality counterfactuals, and our post-processing scheme generates sequential educational instructions to guide learning process. These findings suggest that our method is effective in providing meaningful explanations that connects KT with its educational purpose and priorities. Although our method is sensitive to initialization due to the nature of the KT domain, it opens up new possibilities for handling categorical features in counterfactual explanations. Future work will involve conducting a user study to assess practical impact of KTCF on student learning outcomes and the potential of LLMs to convert counterfactual explanations to educational instructions. Overall, our work contributes to encouraging active discussions on counterfactual XAI for KT and a step towards responsible, stakeholder-centered applications of AI in education.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No.RS-2025-25431740). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(Ministry of Science and ICT)(No.RS-2025-16064585).

References

Afrin, F.; Hamilton, M.; and Thevathyan, C. 2023. Exploring counterfactual explanations for predicting student success. In *International Conference on Computational Science*, 413–420. Springer.

Alfredo, R.; Echeverria, V.; Jin, Y.; Yan, L.; Swiecki, Z.; Gašević, D.; and Martinez-Maldonado, R. 2024. Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6: 100215.

Bai, B.; Liang, J.; Zhang, G.; Li, H.; Bai, K.; and Wang, F. 2021. Why attentions may not be interpretable? In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 25–34.

Bai, Y.; Zhao, J.; Wei, T.; Cai, Q.; and He, L. 2024. A survey of explainable knowledge tracing. *Applied Intelligence*, 54(8): 6483–6514.

Baker, F. B. 2001. *The basics of item response theory*. ERIC.

Block, J. H. 1971. *Mastery learning: Theory and practice*. Holt.

Bloom, B. S. 1968. Learning for mastery. *Evaluation Comment*, 1(2): 1—-12. (ERIC Document Reproduction No. ED053419).

Bloom, B. S. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6): 4–16.

Chen, J.; Liu, Z.; Huang, S.; Liu, Q.; and Luo, W. 2023. Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 14196–14204.

Cui, J.; Yu, M.; Jiang, B.; Zhou, A.; Wang, J.; and Zhang, W. 2024. Interpretable knowledge tracing via response influence-based counterfactual reasoning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 1103–1116. IEEE.

Delaney, E.; Greene, D.; and Keane, M. T. 2021. Instance-based counterfactual explanations for time series classification. In *International conference on case-based reasoning*, 32–47. Springer.

EU. 2024. REGULATION (EU) 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689. Accessed: 2025-07-25.

Freiesleben, T.; and König, G. 2023. Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In *World conference on explainable artificial intelligence*, 48–65. Springer.

Ghosh, A.; Heffernan, N.; and Lan, A. S. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2330–2339.

Guidotti, R. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5): 2770–2824.

Huang, C.-Q.; Huang, Q.-H.; Huang, X.; Wang, H.; Li, M.; Lin, K.-J.; and Chang, Y. 2024. XKT: toward explainable knowledge tracing model with cognitive learning theories for questions of multiple knowledge concepts. *IEEE Transactions on Knowledge and Data Engineering*, 36(11): 7308–7325.

Huang, T.-H.; Metzger, A.; and Pohl, K. 2021. Counterfactual explanations for predictive business process monitoring. In *European, Mediterranean, and Middle Eastern Conference on Information Systems*, 399–413. Springer.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. arXiv:1611.01144.

Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 353–362.

Khosravi, H.; Shum, S. B.; Chen, G.; Conati, C.; Tsai, Y.-S.; Kay, J.; Knight, S.; Martinez-Maldonado, R.; Sadiq, S.; and Gašević, D. 2022. Explainable artificial intelligence in education. *Computers and education: artificial intelligence*, 3: 100074.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Lam, F.; and Newman, A. 2008. Traveling salesman path problems. *Mathematical Programming*, 113(1): 39–59.

Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; and Baum, K. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial intelligence*, 296: 103473.

Liu, Y.; Li, H.; Guo, Y.; Kong, C.; Li, J.; and Wang, S. 2022a. Rethinking attention-model explainability through faithfulness violation test. In *International conference on machine learning*, 13807–13824. PMLR.

Liu, Z.; Guo, T.; Liang, Q.; Hou, M.; Zhan, B.; Tang, J.; Luo, W.; and Weng, J. 2025. Deep Learning Based Knowledge Tracing: A Review, A Tool and Empirical Studies. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; Tang, J.; and Luo, W. 2022b. pyKT: a python library to benchmark deep learning based knowledge tracing models. *Advances in Neural Information Processing Systems*, 35: 18542–18555.

Liu, Z.; Liu, Q.; Guo, T.; Chen, J.; Huang, S.; Zhao, X.; Tang, J.; Luo, W.; and Weng, J. 2023. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems*, 36: 32958–32970.

Lu, Y.; Wang, D.; Chen, P.; Meng, Q.; and Yu, S. 2023. Interpreting deep learning models for knowledge tracing. *International Journal of Artificial Intelligence in Education*, 33(3): 519–542.

Lu, Y.; Wang, D.; Chen, P.; and Zhang, Z. 2024. Design and evaluation of trustworthy knowledge tracing model for intelligent tutoring system. *IEEE Transactions on Learning Technologies*, 17: 1661–1676.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Moreira, C.; Chou, Y.-L.; Hsieh, C.; Ouyang, C.; Pereira, J.; and Jorge, J. 2025. Benchmarking instance-centric counterfactual algorithms for XAI: from white box to black box. *ACM Computing Surveys*, 57(6): 1–37.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 607–617.

Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44): 22071–22080.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; Van Keulen, M.; and Seifert, C. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s): 1–42.

Nguyen, A.; Ngo, H. N.; Hong, Y.; Dang, B.; and Nguyen, B.-P. T. 2023. Ethical principles for artificial intelligence in education. *Education and information technologies*, 28(4): 4221–4241.

Office of Educational Technology. 2023. Artificial intelligence and the future of teaching and learning: Insights and recommendations. *U.S. Department of Education, Washington, DC*.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Qin, Y.; Zhu, X.; Tang, X.; Zhang, C.; Wu, K.; Chang, F.; Diao, J.; and Hu, Z. 2025. Interpretable Knowledge Tracing with Difficulty-Aware Attention and Selective State Space Model. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 316–325.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Association for Computational Linguistics.

Smith, B. I.; Chimedza, C.; and Bührmann, J. H. 2022. Individualized help for at-risk students using model-agnostic and counterfactual explanations. *Education and Information Technologies*, 27(2): 1539–1558.

Sun, J.; Yu, F.; Wan, Q.; Li, Q.; Liu, S.; and Shen, X. 2024. Interpretable knowledge tracing with multiscale state representation. In *Proceedings of the ACM Web Conference 2024*, 3265–3276.

Tsiakmaki, M.; and Ragos, O. 2021. A case study of interpretable counterfactual explanations for the task of predicting student academic performance. In *2021 25th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, 120–125. IEEE.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, 10–19.

Valero-Leal, E.; Carlon, M. K. J.; and Cross, J. S. 2023. A shap-inspired method for computing interaction contribution in deep knowledge tracing. In *International conference on artificial intelligence in education*, 460–465. Springer.

Van Looveren, A.; and Klaise, J. 2021. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 650–665. Springer.

Verma, S.; Boonsanong, V.; Hoang, M.; Hines, K.; Dickerson, J.; and Shah, C. 2024. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12): 1–42.

Vincent-Lancrin, S.; and Van der Vlies, R. 2020. Trustworthy artificial intelligence (AI) in education: Promises and challenges. *OECD education working papers*, (218): 1–17.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wang, D.; Chen, Z.; Florescu, I.; and Wen, B. 2023a. A sparsity algorithm for finding optimal counterfactual explanations: Application to corporate credit rating. *Research in International Business and Finance*, 64: 101869.

Wang, D.; Lu, Y.; Zhang, Z.; and Chen, P. 2022. A generic interpreting method for knowledge tracing models. In *International conference on artificial intelligence in education*, 573–580. Springer.

Wang, X.; Zhao, S.; Guo, L.; Zhu, L.; Cui, C.; and Xu, L. 2023b. GraphCA: Learning from graph counterfactual augmentation for knowledge tracing. *IEEE/CAA Journal of Automatica Sinica*, 10(11): 2108–2123.

Wang, Z.; Samsten, I.; and Papapetrou, P. 2021. Counterfactual explanations for survival prediction of cardiovascular ICU patients. In *International conference on artificial intelligence in medicine*, 338–348. Springer.

Warren, G.; Byrne, R. M.; and Keane, M. T. 2024. Categorical and continuous features in counterfactual explanations of AI systems. *ACM Transactions on Interactive Intelligent Systems*, 14(4): 1–37.

Zhang, H.; Dong, J.; Lv, C.; Lin, Y.; and Bai, J. 2023a. Visual analytics of potential dropout behavior patterns in online learning based on counterfactual explanation. *Journal of Visualization*, 26(3): 723–741.

Zhang, M.; Zhu, X.; Zhang, C.; Qian, W.; Pan, F.; and Zhao, H. 2023b. Counterfactual Monotonic Knowledge Tracing for Assessing Students' Dynamic Mastery of Knowledge Concepts. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, 3236–3246.

Zhao, J.; Bhatt, S.; Thille, C.; Zimmaro, D.; and Gattani, N. 2020. Interpretable personalized knowledge tracing and next learning activity recommendation. In *Proceedings of the seventh ACM conference on learning@ scale*, 325–328.