

RESEARCH

Free and Open Access

What's more important when developing math recommender systems: accuracy, explainability, or both?

Yiling Dai ¹*, Brendan Flanagan ² and Hiroaki Ogata ¹

*Correspondence:
dai.yiling.4t@kyoto-u.ac.jp
Academic Center for Computing
and Media Studies,
Kyoto University,
Kyoto, Japan
Full list of author information is
available at the end of the article

Abstract

To make accurate predictions, complex artificial intelligence techniques are being adopted in intelligent systems. It leads to the need for explanations, helping users understand how the model works. Beyond this original purpose, explanations in educational intelligent systems have been found to increase students' awareness, perceived usefulness, and acceptance of the recommendations. Can we ensure a model makes accurate predictions and has effects of explanations at the same time? Though it is commonly considered that complex models are accurate but difficult to interpret, it remains debatable whether there is a trade-off between the accuracy and explainability of such models. In this study, we explore the relationships between accuracy and explainability of different models for recommending math quizzes in the context of formative assessment. Focusing on three recommender models—an inherently explainable model (Naïve CE), a black-box model (MF), and an integrated model (CE+MF), we compared the accuracy using a large-scale real-world dataset and evaluated the explanations in a semi-interactive questionnaire survey. We found that: 1) There was a trade-off between accuracy and explainability given the specific context. 2) The explainability did not demonstrate consistent trends among different aspects. Especially, perceived understandability did not indicate the perceived usefulness in math learning and the behavioral intention to use the system. 3) The integrated model displayed a balanced level of accuracy and explainability, which implies the feasibility to develop an explainable educational recommender system by improving the accuracy of an inherently explainable model.

Keywords: Recommender systems, Math quizzes, Explainability, Accuracy, Matrix factorization



© The Author(s). 2025 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Introduction

Artificial intelligence (AI) techniques have been adopted in various intelligent systems to empower the task performance (Adadi & Berrada, 2018; Arrieta et al., 2020; Vultureanu-Albiși & Bădică, 2022). Unlike traditional machine learning techniques, recent models have become more complex and bring about the need to explain how it works for humans (Khosravi et al., 2022; Vilone & Longo, 2021). Aspects of the explanation such as justification, effectiveness, efficiency, informativeness, and persuasiveness have been researched beyond the original motivation (Adadi & Berrada, 2018; Tintarev & Masthoff, 2007; Vilone & Longo, 2021). Explanations in educational intelligent systems have been found to increase students' awareness, perceived usefulness, acceptance of the recommended items (Barria-Pineda et al., 2021; Conati et al., 2021; Dai, Takami, et al., 2022; Hur et al., 2022; Takami et al., 2022). It is important and beneficial to provide explanations when adopting advanced techniques in intelligent systems to make an impact on learning.

Can we ensure a model that makes accurate predictions and has effects of explanations at the same time? Intuitively, more complex models have better performance at making predictions but are more difficult to interpret. However, deep models do not always outperform simpler ones, and complexity does not necessarily reduce explainability (Bell et al., 2022; Gervet et al., 2020). In addition, it is challenging to define and measure explainability consistently (Vilone & Longo, 2021). Consequently, the accuracy-explainability trade-off is rarely verified in practice (Bell et al., 2022; Molnar et al., 2022; Rudin, 2019). Given the contrasting arguments in the discussion of the trade-off between accuracy and explainability and the difficulty to measure explainability, it is necessary to explore it in the specific context of our interest.

If there is a trade-off between the accuracy and explainability of a model, what is the better approach to develop an explainable accurate model? Should we develop an explainable model in the first place and then improve the accuracy or develop a complex model and try to explain it afterwards (Molnar et al., 2022)? Over-using complex models was concerned while an inherently explainable model which has a comparable accuracy is available (Khosravi et al., 2022; Molnar et al., 2022). To address the forementioned issues, we focus on a formative assessment context, where the system estimates mastery levels and recommends math quizzes for K-12 students. By comparing three recommender models—an inherently explainable model, a black-box model, and an integrated model, our goal is to answer the following questions:

RQ1: Is there a trade-off between the accuracy and the explainability of recommender models in the context of K-12 math learning?

RQ2: Is it feasible to enhance the accuracy of inherently explainable educational recommender systems by integrating them with black-box models?

Specifically, we focused on a recommender system named Naïve Concept Explicit (Naïve CE) (Dai, Flanagan, et al., 2022) which recommends quizzes based on the estimation of the students' mastery level on math concepts and provides explanations alongside with the recommendations. We selected Naïve CE as it serves as an example of an inherently explainable model which adopts shallow computations with human-readable math concepts. Besides, Naïve CE has a comparable black-box model—Matrix Factorization (MF), which is commonly used in general recommender systems. We also proposed an integrated model CE+MF with the expectation of preserving the characteristics of both models in terms of accuracy and explainability. We then compared the accuracy of three models using a large-scale real-world dataset of student quiz answers. For the explainability, we conducted a pilot study to explore how to generate and evaluate the explanations. We designed a semi-interactive questionnaire and collected the opinions of 12 participants on different aspects of the explanations. We found that: 1) There was a trade-off between the accuracy and explainability of the models given the specific context of a math recommender system. 2) The explainability demonstrated different trends among perceived understandability, perceived usefulness in math learning and the behavioral intention to use the system. 3) The integrated model displayed a balanced level of accuracy and explainability.

The contribution of this study is threefold:

- We evaluated and compared the accuracy and explainability of educational recommender systems under realistic settings.
- We explored the way to measure explainability of recommender systems from educational perspectives, which is an important step towards the enrichment of explainable AI research.
- Our approach served as an example to develop explainable recommender systems by integrating inherently explainable models with black-box models.

The remaining part of this paper is structured as follows. We first review and summarize the related works in both general and educational contexts. We then describe the basics of three models to be compared. Afterwards, we evaluate the accuracy and explainability of the models, respectively. Lastly, we present results, discussion, conclusion and future work.

Related works

Explanations in intelligent systems

The discussion about explainable artificial intelligence (XAI) stems from the lack of transparency in black-box models (Adadi & Berrada, 2018). For example, recent AI techniques such as neural networks involve complex structures of hidden layers, which are

difficult to understand for humans (Khosravi et al., 2022; Vilone & Longo, 2021; Vultureanu-Albiși & Bădică, 2022). Early machine learning methods such as decision trees and linear regression are considered as transparent as the outcomes are traceable and easy-to-understand to humans (Arrieta et al., 2020). With more attention from the researchers and practitioners, the purposes and effects of explanations have been extended to justification, control, improvement, effectiveness, efficiency, informativeness, persuasiveness, trust, satisfaction, and so on (Adadi & Berrada, 2018; Tintarev & Masthoff, 2007; Vilone & Longo, 2021; Vultureanu-Albiși & Bădică, 2022).

Due to the diverse purposes of providing explanations, there is still no agreement on the definition of explainability among scholars (Adadi & Berrada, 2018; Vilone & Longo, 2021). In this study, we follow Miller's (2019) notion of "explainability", which is the ability of a system making the human-user to understand its decisions during the computation.

Discussions about accuracy and explainability

Complex black-box models are commonly considered to be more accurate but less interpretable. However, this trade-off between accuracy and explainability is rarely confirmed in practice, with conflicting evidence found in research (Arrieta et al., 2020; Bell et al., 2022; Guleria & Sood, 2023; Molnar et al., 2022; Rudin, 2019; Vilone & Longo, 2021). Deep knowledge tracing models did not always generate better predictions than logistic regression models as the size and shape of dataset changed (Gervet et al., 2020). Guleria and Sood (2023) compared the performance of typical white-box and black-box classifiers for predicting job placement. They found that naïve Bayes models outperformed black-box models such as ensemble models and neural network. It is difficult to quantify explainability compared with the richness of metrics to measure accuracy (Bell et al., 2022). This motivates us to explore whether there is a trade-off between the accuracy and the explainability of recommender systems in the specific context of math learning.

Given the uncertainty of the trade-off between accuracy and explainability, another intertwined question is how to develop an accurate and explainable model. Basically, model-intrinsic explanations can be generated easily from the model itself if the model is inherently explainable. Post-hoc explanations can be added if the model is complex in its nature (Zhang & Chen, 2020). In educational contexts, an example of model-intrinsic explanation can be to explain how the student's knowledge state is estimated and why a learning item is considered preferable to improve his/her knowledge state (Dai, Flanagan, et al., 2022). In contrast, a post-hoc explanation for a recommended item can be something not necessarily related to the knowledge state estimation but instrumental in motivating the student to accept the recommendation. For instance, an explanation showing how many

students have attempted this item may work for students who are weak to peer pressure (Takami et al., 2023).

Adopting post-hoc approaches could be risky for high-stakes domains as the explanations may provide misleading information (Rudin, 2019). For instance, feature-based explanations were adopted for black-box models but Swamy et al. (2022) found that the explainers are not consistent on feature importance. If an inherently explainable model has a comparable accuracy, over-using complex models is not recommended (Khosravi et al., 2022; Molnar et al., 2022). What is the better way to develop accurate and explainable models? In this study, we investigate the accuracy and explainability of an inherently explainable model, a complex model, and an integrated model. Generating explanations for the complex model serves as the approach to explain a complex model and the integrated model serves as the approach to improve the accuracy of an explainable model.

Types and principles of explanations

Notions and characteristics of explainability should be expanded based on the domains in which it is applied (Vilone & Longo, 2021). From cognitive perspective, explanation is fundamental to human's sense of understanding and has profound effects on causal inference and learning (Lombrozo, 2006). Therefore, two key elements of developing explainable educational recommender systems are—1) what explanations help improve students' learning performance, 2) and how students react and perceive the explanations.

Among various principles of generating explanations, a pair of contradictory principles is being complete and not overwhelming (Kulesza et al., 2015). Similarly, Ribeiro et al. (2016) named it as a trade-off between interpretability and fidelity of the explanations. A faithful explanation should describe the model completely to fulfill the fidelity. In contrast, an interpretable explanation should take human limitations into account. Following this argument, we divide explanations of intelligent systems into model-oriented and user-oriented. For inherently explainable models, a complete explanation tends to be interpretable to users by its nature. Rule-based explanations are commonly adopted for models such as decision tree, fuzzy rules, and association rules (Alonso & Casalino, 2019; Conati et al., 2021). For black-box models, feature-based explanations are model-oriented as it uses “model language”. For example, Swamy et al. (2022) implemented different feature-based explainers in the neural network model of predicting student performance. They utilized the explainers to investigate the performance as model developers, who have knowledge to understand information such as feature importance scores. While knowledge-based explanations emphasize how end users can process and utilize the information for decision making. For example, Wang et al. (2018) extracted properties such as “neat”, “vegan”, and “sandwich”, which help users decide whether to accept a recommended restaurant.

In educational context, previous works investigated effects of explanations such as increasing students' awareness, trust, and intention (Barria-Pineda et al., 2021; Conati et al., 2021; Ooge et al., 2022; Takami et al., 2023; Yu et al., 2021). However, the understandability and the fidelity of explanations, and how they interplay with other learning effects is underexplored. In this study, we aim at investigating whether the explanations a) help students understand how the model works, b) help math learning, and c) motivate students to adopt the recommendation in models of different types and performance.

Preliminaries

Naïve CE for math recommendations

Formative assessment implemented with individualized tools is considered beneficial to students (Faber et al., 2017). As one type of digital formative assessment activity, we focus on the learning scenario where the students practice math quizzes and get feedback from a recommender system. Estimating the probability of a student mastering a given skill, which is called cognitive diagnosis, is an essential step to provide adaptive feedback (Desmarais & Pelczar, 2010). Rather than the final scores, namely, the probability of correctly answering a quiz, identifying specific missing concepts and difficult areas is more instructive (Birenbaum et al., 1993). As a result, it is desirable that the model recommends quizzes and estimates student skill levels by readable math concepts.

Models such as item response theory (Yen & Fitzpatrick, 2006) and knowledge tracing (Corbett & Anderson, 1994), predict students answers by modeling it with parameters such as student skill level, item difficulty, knowledge states, learning rates, and so on. These models can estimate mastery levels of predefined skills or concepts, but usually involve complex probabilistic reasoning in the estimation. In this study, we select a concept-explicit recommender system Naïve CE (Dai, Flanagan, et al., 2022) as an instance of inherently explainable model, where students' probabilities of mastering math concepts are estimated in a shallow and human-understandable manner.

However, we also have a concern on the estimation performance of Naïve CE. In other words, how well can Naïve CE estimate students' mastery level and their probabilities to correctly answer the quizzes? Is there an inferiority in estimation performance compared with more accurate but less explainable models? As Barnes (2005) pointed out, it is debatable whether an explicit model with expert-assigned concepts models student performance better than an implicit model with latent factors. Therefore, we selected a classic recommendation model matrix factorization with latent factors as a comparative target of Naïve CE.

MF for general and math recommendations

Matrix factorization (MF) (Koren et al., 2009; Takács et al., 2008) is a frequently used model to recommend items that users may have interests. This model assumes that a user's interest in an item comes from her/his preferences on some factors and the relatedness of the factors with the item. It then guesses the unseen user-item interactions by learning from observed user-item interactions. Khosravi et al. (2017) applied MF in their model to estimate students' knowledge gaps to answering the quizzes. To integrate the strengths of both models, Abdi et al. (2018) fed the error in Bayesian knowledge tracing model to an MF model, improving the accuracy of estimating student performance. Since MF has been widely adopted and evaluated as a useful model to estimate student performance and shares some similarities in modeling the problem with Naïve CE, we chose MF as a comparative model in discussing the estimation performance and model explainability. We also propose a method to integrate two models so that the readability of concepts in Naïve CE is preserved.

Math recommender models

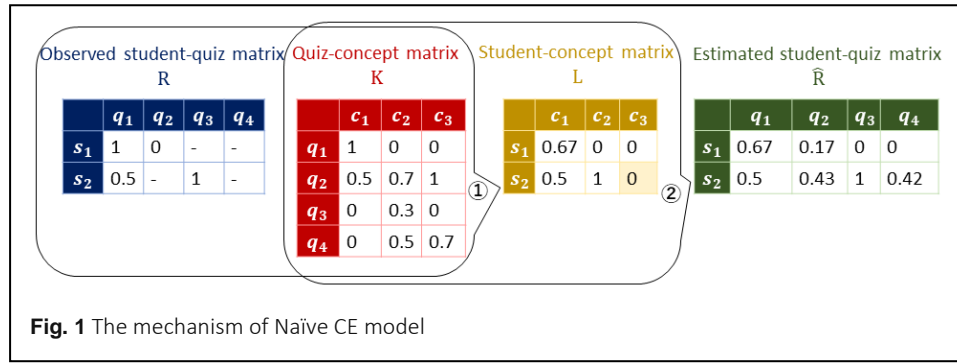
Problem definition

In the recommender systems, the learning activity is modeled as a sequence of students' reactions towards quizzes. The task is to recommend quizzes that fit an individual student's learning progress. It is common that the observed student reactions are limited to a small set of quizzes. As a result, predicting student reactions on unseen quizzes is a key step. We formalize the problem as follows: Given a set of m students, a set of n math quizzes, and the observed student correctness rates on the quizzes $R \in \mathbb{R}^{m \times n}$ (whose entry r_{ij} indicates the correctness rate of quiz q_j by student s_i), we want to estimate the student correctness rates on the whole set of quizzes $\hat{R} \in \mathbb{R}^{m \times n}$.

Naive Concept-Explicit Model (Naïve CE)

In this section, we briefly review the basic idea and mechanism of Naïve CE, which is mainly regenerated from Dai, Flanagan, et al.'s work (2022). Suppose that we have a quiz "Let the set of all positive divisors of 12 be A . Fill in the \square with \in or \notin . (1) $2 \square A$ (2) $7 \square A$ (3) $12 \square A$ ". Solving the math quiz requires the knowledge of "set" and "positive divisor". The probability that a student can successfully solve a math quiz depends on how s/he understands the required concepts.

STEP 1 We are given the observed student-quiz matrix R whose entry r_{ij} indicates the correctness rate of quiz q_j by student s_i , and the quiz-concept matrix K whose entry k_{jo} indicates the relatedness of a quiz q_j and a concept c_o , we calculate the students' concept



mastery level L by looking at how they successfully solved quizzes related to the concept. The calculation can be abstracted as $L = R \cdot K$. Note that the quiz-concept relatedness is extracted from the quiz information automatically, which is also readable concepts to students.

STEP 2 We then estimate the probability of a student successfully solving a quiz by considering how many of the required concepts have been mastered. The calculation can be abstracted as $\hat{R} = L \cdot K^T$.

By doing this, the probabilities are modified by the inter-relationships between quizzes and concepts. For instance, s_1 successfully solved q_1 in history but got an estimated success of 0.67. This is because q_1 requires the knowledge of c_1 and the student failed to solve q_2 which also requires the knowledge of c_1 . However, this model falls short in coping with unseen concepts. For instance, s_2 had not attempted any quizzes related to c_3 . As a result, c_3 is ignored in STEP 2.

Matrix Factorization (MF)

We apply MF in the setting of solving math quizzes. MF decomposes the observed student-quiz interaction matrix $R \in \mathbb{R}^{m \times n}$ into two matrices $P \in \mathbb{R}^{m \times t}$ and $F \in \mathbb{R}^{t \times n}$ such that $R \approx PF$, where t is the number of latent factors. P is the student-factor matrix whose entry p_{it} represents the mastery level of factor f_t by student s_i . F is the factor-quiz matrix whose entry f_{tj} represents the relatedness of factor f_t and quiz q_j . By minimizing the difference between the estimated and the observed interactions (also viewed as a machine learning process), we get full P and F , which help us estimate the unseen interactions. This underlying idea is somehow similar to Naïve CE except that the factors are “latent” and difficult to interpret. There is a variant of MF which considers user bias and item bias. As a user may tend to highly rate all items or an item of low quality tends to be rated low by all users, introducing bias parameters in MF helps model this situation. In this setting, user bias and item bias can be interpreted as student ability and quiz difficulty. By doing this,

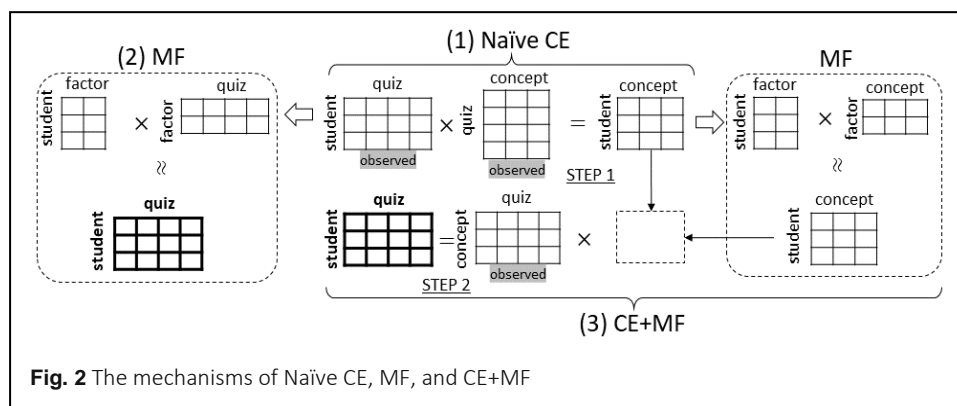
the sum of PF and bias better approximates the observed interactions but the intermediate value of PF is harder to interpret. We omit mathematical details of the bias model for the sake of conciseness.

Concept-Explicit Matrix Factorization (CE+MF)

Previously, we discussed that Naïve CE is easy to interpret since the concepts are predefined and the computation is simple and straightforward. However, it simply gives up guessing when encountering unseen concepts and quizzes. In contrast, MF is good at estimating the probable values for unseen interactions by iteratively learning from the observed interactions. However, the latent factors of the consisting matrices are difficult to interpret. To take advantage of both models, we propose a simple hybrid model called CE+MF. As illustrated in Figure 2, Naïve CE utilized the observed student-quiz matrix and quiz-concept matrix to estimate student-concept matrix. The student-concept matrix is again used to adjust the student-quiz matrix. MF simply decomposes the observed student-quiz matrix into two matrices with latent factors. In CE+MF, we first estimate the student-concept matrix as we do in STEP 1 in Naïve CE model. Then, we adopt MF model to update the student-concept matrix where the mastery level on unseen concepts is modified. Last, we update the student-quiz matrix with the updated student-concept matrix as we do in STEP 2 in Naïve CE model. As a result, CE+MF model is supposed to have a higher predictive performance than Naïve CE model while preserving the explainability on concepts. To avoid redundancy, we omit the mathematical details of CE+MF.

Evaluation

In this study, we aim at investigating two aspects of the recommender models—accuracy and explainability. For accuracy, we capture it as the quiz mastery level estimation performance, and use a historical dataset collected in a learning system to evaluate whether the models can correctly predict students' answers for unseen quizzes. For explainability,



we conduct a questionnaire survey where the explanations are displayed in a semi-interactive manner. We then collect participants' opinions of explainability from different perspectives.

Accuracy: Quiz mastery level estimation performance

Dataset

We collected quiz answering data from a learning system (Flanagan et al., 2021) of the first-year students in a Japanese high school from April 2021 to March 2022. During this period, the students attempted the math quizzes in different contexts such as finishing the assignments, preparing for an upcoming test, and self-oriented practicing. As a part of the adjustment and correction in the formative assessment, they were required to check the answer and report whether they solved the quiz successfully after attempting a quiz. The system language and quiz language were Japanese. Each attempt was recorded as a 0-1 score associated with the student id, quiz id, and timestamp. We computed the aggregated student-quiz correctness rate by taking the average score of all attempts throughout the period. We did not conduct any data filtering process as the temporal order of attempts and the number of attempts are not essential in this evaluation framework. Finally, we obtained a dataset consisting of 27,431 attempts for 270 unique students and 1,919 unique quizzes. Table 1 shows the statistics of the number of attempts per student and per quiz. After converting the log data into the student-quiz correctness matrix, only 23,155 pairs of students and quizzes were observed, which indicates a very high sparsity of 95.53%

$$\left(1 - \frac{\# \text{ observed pairs}}{\# \text{ students} \cdot \# \text{ quizzes}}\right).$$

Metrics

Our main concern is to evaluate whether a model can predict a student's success probability in a quiz. Therefore, we adopt two metrics to measure the agreement between the estimated probability and true correctness rates: **Area under ROC curve (AUC)** is considered an effective metric to measure how well a model separate negative and positive samples across different decision threshold choices (Bradley, 1997). Since the true correctness rates for student-quiz pairs are real numbers between 0 to 1, we first transform the true correctness

Table 1 Statistics of the dataset for quiz mastery level estimation evaluation

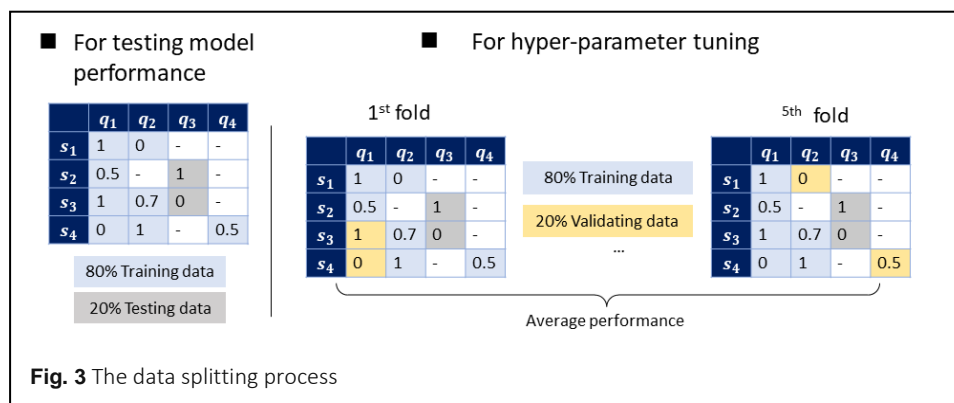
	# attempts per student	# attempts per quiz
Mean	101.596	14.294
SD	97.681	31.404
Min	1	1
Max	808	426

rates into 1 if it is greater than 0.5, 0 otherwise when applying AUC. **Root mean square error (RMSE)** is used to measure the absolute differences between the estimated probability and the true correctness rates.

Implementation

As illustrated in the left part of Figure 3, we randomly set aside 20% of the student-quiz correctness rate values as test data and all models were blind to these data during training or tuning process. For models involving hyper-parameter tuning, we adopted a 5-fold cross validation approach to select the best combination of parameters. As illustrated in the right part of Figure 3, in each fold, 20% of the data is used to validate the model's performance and the average performance of all the folds is treated as the final performance of a combination of parameters. We adopted a grid-search approach to generate combinations of parameters. Only the performance of the best combination of parameters will be reported in the following section.

The following are some implementation details of the models: 1) Naïve CE. As described in Dai, Flanagan, et al. (2022), we adopted text mining techniques to automatically extract math concepts from the quizzes. The entries of quiz-concept matrix were computed using term frequency-inverse document frequency weighting scheme (Salton & Buckley, 1988). 2) MF. We adopted stochastic gradient descent algorithm to obtain the matrices P and F whose product has the minimum difference with the observed student-quiz correctness rate. We tuned three parameters—learning rate α , regularization factor γ and number of latent factors k for MF. MF_bias and CE+MF_bias are variants with bias parameters of MF and CE+MF, respectively. The best combination of hyperparameters is $\alpha = 0.01$, $\gamma = 0.1$, $k = 5$ and is used in all MF-related models. The code for MF-related parts was adapted from Yeung (2020).



Results

Table 2 shows the AUC and RMSE values for each model. Overall, Naïve CE has the lowest, MF has the highest, and CE+MF has the medium performance in both metrics. MF with bias has better performance both in separate and hybrid models. This result is consistent with our expectation:

- Naïve CE is straightforward but ignores unseen concepts or quizzes. From the perspective of AUC, this model can discriminate between correct or incorrect answers. However, the performance was nearly a random model from the perspective of RMSE, which means the detailed values of the correctness has a large gap with student's true mastery level.
- MF and MF_bias did a good job at approximating the observed student-quiz correctness rates and therefore the latent factors helped to predict the values for unseen pairs. With the AUC value being close to 0.8, MF_bias is supposed to be practically useful to separate a correct or incorrect answer (Mandrekar, 2010). Meanwhile, the RMSE value was still high if we consider a situation where we mistake a student's correctness rate 0.98 into 0.6. However, whether the students can recognize the difference and how they perceive the estimation needs to be further investigated and discussed.
- The hybrid model CE+MF achieved better performance than Naïve CE but still has a distance to MF. We consider a possible reason to be the flaws in quiz-concept matrix. First, not all necessary knowledge and skills for solving a math quiz can be detected from the textual information of the quiz. Second, the relatedness of the concepts to a quiz may not be correct just judging from their occurrences in the quiz. Ingesting more elaborate domain models to Naïve CE and observing the performance improvement is one of the future directions.

Explainability

Before evaluating with real users at a large-scale, we conducted a preliminary survey to reduce the uncertainty of the measurement of explainability. The survey is simulation-based, where the participants do not really answer quizzes but only evaluate recommendations with synthesized answering history. Admittedly, the setting might be

Table 2 Quiz mastery level estimation evaluation results

Model	AUC	RMSE
Naïve CE	0.639	0.508
CE+MF	0.688	0.478
CE+MF_bias	0.692	0.464
MF	0.772	0.419
MF_bias	0.799	0.381

limited as the experience of participants is not complete. Fortunately, simulation-based user studies were able to provide insights before real-world deployment (Wang et al., 2018). The following sections present the questionnaire design, survey procedure, and the results, respectively.

Questionnaire design

Following Miller's (2019) notions, we call the statements which explains how the system works as explanations, the agent who generates the explanations as explainer, the human-users who receives the explanations as explainees. The behavior of explaining is an interactive process between the explainer and the explainee (Miller, 2019). Intuitively, the explainee can question any elements s/he does not understand in a statement the explainer makes, which leads to the next explanation. Given this premise, it is intuitive to develop a conversational system between the recommender model and the user. However, conducting unstructured surveys such as interviews is difficult to quantify. In this study, we designed a semi-interactive questionnaire where all the available explanations are prepared in a sequence, and the participants can view them in an interactive manner. After the participants viewed the explanations, we used a 5-point Likert scale to evaluate the perceived understandability, perceived usefulness in math learning, and behavioral intention as listed in Table 3. Single-item measures were used in this study, and we leave well-constructed measures in the future work.

Figure 4 illustrates the overall design of the questionnaire. The participant is supposed to view the explanation sequences and evaluate the models one by one. For each model, the explanations are displayed one by one, and the participant can stop viewing the remaining explanations at any time if s/he is satisfied with the previous explanations. As a result, the participant may give evaluations based on different numbers of explanations for the recommender models. Tables 4 to 6 list the explanation sequences for Naïve CE, MF, and CE+MF, respectively. Naïve CE, MF, and CE+MF are versatile at making various mathematical estimations. In this study, we focus on exploring their explainability on a common estimation. Therefore, the explanations of all three models start with the same recommendation of a specific mathematical quiz, followed with two explanations which

Table 3 Items to evaluate explainability

Aspect	Item	Source
Perceived understandability	I understand how the system works and makes its judgements.	Chatti et al. (2022)
Perceived usefulness in math learning	I think the system and its explanations are helpful in my math learning.	Original
Behavioral intention to use the system	I would like to use the system in real settings.	Tintarev & Masthoff (2007)

indicate the estimated difficulty of the quiz. The explanations diverge from the third statements based on the algorithms behind the models. Note that we randomly sampled a quiz and synthesized the values to avoid bias. For instance, the explanations of Naïve CE then clarify how the difficulty is estimated by introducing the extracted mathematical concepts, their relevance with the quiz, how the mastery level is estimated from the answering history. Finally, the explanations end with an overview of the algorithm with some simple annotations. The explanations for MF and CE+MF are generated following the same policy. For MF, the explanations are the shortest as the latent factors are not explainable and the rationale of matrix factorization is too complex to be explained. In other words, many elements of the algorithm in MF model are still encapsulated in the explanations. For CE+MF, the explanations are the longest as it integrates Naïve CE with MF in terms of the algorithm. Note that all the explanations are generated from the perspective of explaining how a recommendation is made in the system in a student-friendly manner. When we go deeper into the explanation sequence, we expect that it may not be of students' interests to know the very details of the system. However, it is meaningful to include these explanations in a preliminary evaluation to explore participants' opinions.

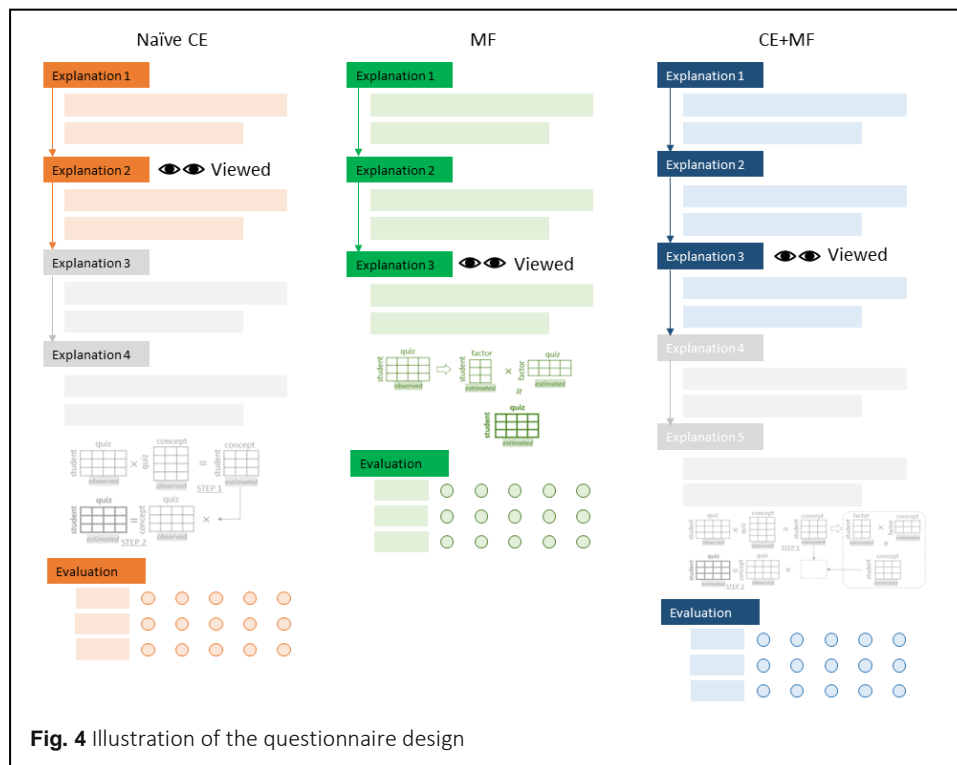


Fig. 4 Illustration of the questionnaire design

Table 4 Explanation sequence of Naïve CE model

Number	Content
Recommendation	Quiz 129: The area decided inequality (boarder line is a parabola) Plot the area of the inequality in the graph. (1) $y > x^2 + 2x$ (2) $y \leq 2x^2 - 8x + 9$
Explanation 1	I recommend this quiz as I think it is challenging for you.
Explanation 2	The estimated probability of you answering this quiz correctly is 0.2002. Since the probability is quite low, I think it is challenging for you.
Explanation 3	I estimate the probability based on your understanding of the math concepts [inequality (不等式), parabola (放物線)] that are necessary to solve the quiz. Specifically, I estimate your level of understanding of inequality is 0.28, and your understanding of parabola is 0.18.
Explanation 4	[inequality (不等式), parabola (放物線)] are considered to be necessary math concepts as they appear in the quiz or the standard solution.
Explanation 5	Your level of understanding of inequality is estimated from your attempts on other quizzes which also requires the knowledge of inequality. Specifically, you have got Quiz 120 wrong once (correctness rate=0/1=0.0), and you attempted Quiz 121 twice and got it right for the second time (correctness rate=1/2=0.5). Then, I estimate your level of understanding of inequality by taking the weighted average of the correctness rates.
Explanation 6	The weight of math concepts to a quiz is decided from the term-weighting perspective. If a math concept appears frequently in a quiz and not in other quizzes, it is considered to be important to this quiz. Therefore, inequality has different weights in Quiz 120 and Quiz 121, which affects the estimated level of understanding.
Explanation 7	The following illustration summarizes how I work. student-quiz matrix (observed): you and other students' correctness rates on the quizzes. quiz-concept matrix (observed): concepts' weights in the quizzes. student-concept matrix (estimated): estimated levels of understanding of the concepts. student-quiz matrix (estimated): probability of answering quizzes correctly.

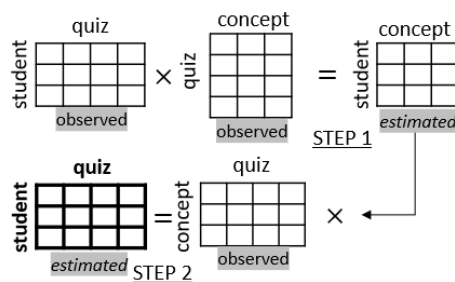


Table 5 Explanation sequence of MF model

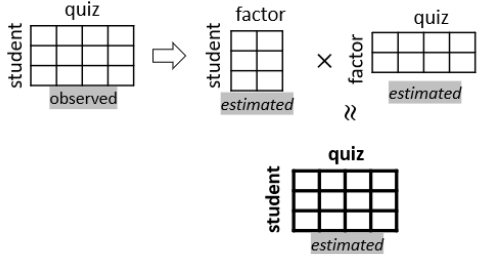
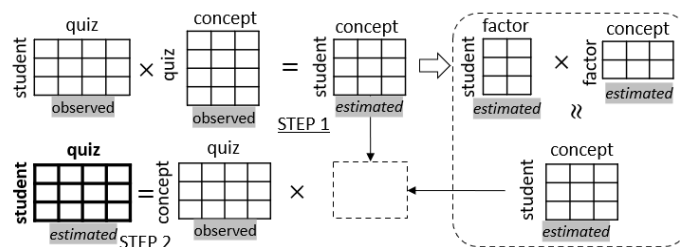
Number	Content
Recommendation	Same as Naïve CE model.
Explanation 1	
Explanation 2	
Explanation 3	I estimate the probability based on your understanding of some latent math factors that are necessary to solve the quiz. Let's call them factor 1 and factor 2. (Unfortunately, I don't know what exactly they are.) Then, I suppose [correctness rate of answering Quiz 120 = weight of factor 1 in Quiz 120 * understanding level of factor 1 + weight of factor 2 in Quiz 120 * understanding level of factor 2.]
Explanation 4	Since I can observe your and other students' correctness rates of answering the quizzes from your attempts, I can guess the values of weights of the factors and understanding levels of the factors through a mathematical method called Matrix Factorization.
Explanation 5	<p>The following illustration summarizes how I work.</p> <p>student-quiz matrix (observed): you and other students' correctness rates of the quizzes.</p> <p>student-factor matrix (estimated): estimated levels of understanding of the factors.</p> <p>factor-quiz matrix (estimated): estimated weights of factors in quizzes.</p> <p>student-quiz matrix (estimated): probability of answering quizzes correctly.</p>
 <p>The diagram illustrates the Matrix Factorization process. It starts with an observed student-quiz matrix (a 4x4 grid labeled 'observed' with 'student' on the y-axis and 'quiz' on the x-axis). An arrow points to the decomposition into two matrices: an estimated student-factor matrix (a 4x2 grid labeled 'estimated' with 'student' on the y-axis and 'factor' on the x-axis) and an estimated factor-quiz matrix (a 2x4 grid labeled 'estimated' with 'factor' on the y-axis and 'quiz' on the x-axis). These two matrices are multiplied (indicated by a large 'X' and a double arrow '⇒') to produce the final estimated student-quiz matrix (a 4x4 grid labeled 'estimated' with 'student' on the y-axis and 'quiz' on the x-axis).</p>	

Table 6 Explanation sequence of CE+MF model

Number	Content
Recommendation	Same as Naïve CE model.
Explanation 1	
Explanation 2	
Explanation 3	I estimate the probability based on your understanding of the math concepts [inequality (不等式), parabola (放物線)] that are necessary to solve the quiz. Specifically, I estimate your level of understanding of inequality is 0.28, and your understanding of parabola is 0.18.
Explanation 4	[inequality (不等式), parabola (放物線)] are considered to be necessary math concepts as they appear in the quiz or the standard solution.
Explanation 5	Your level of understanding of inequality is estimated from your attempts on other quizzes which also requires the knowledge of inequality. Specifically, you have got Quiz 120 wrong once (correctness rate=0/1=0.0), and you attempted Quiz 121 twice and got it right for the second time (correctness rate=1/2=0.5). Then, I estimate your level of understanding of inequality by taking the weighted average of the correctness rates.
Explanation 6	The weight of math concepts to a quiz is decided from the term-weighting perspective. If a math concept appears frequently in a quiz and not in other quizzes, it is considered to be important to this quiz. Therefore, inequality has different weights in Quiz 120 and Quiz 121, which affects the estimated level of understanding.
Explanation 7	Since you may have only attempted a small number of quizzes, there is not sufficient data to make a good estimation. I also use other students' answers to improve my estimation of your levels of understanding of the concepts.
Explanation 8	I estimate the understanding level of concepts based on some latent math skills that are necessary to solve the quiz. Let's call them factor 1 and factor 2. (Unfortunately, I don't know what exactly they are.) Then, I suppose [understanding of inequality = weight of factor 1 in inequality * level of factor 1 + weight of factor 2 in inequality * level of factor 2.]
Explanation 9	Since I have estimated your and other students' levels of understanding of inequality, I can guess the values of weights of the factors and levels of the factors through a mathematical method called Matrix Factorization.
Explanation 10	<p>The following illustration summarizes how I work.</p> <p>student-quiz matrix (observed): you and other students' correctness rates on the quizzes.</p> <p>quiz-concept matrix (observed): concepts' weights in the quizzes.</p> <p>student-concept matrix (estimated): estimated levels of understanding of the concepts.</p> <p>student-factor matrix (estimated): estimated levels of factors.</p> <p>factor-concept matrix (estimated): estimated weights of factors in concepts.</p> <p>student-quiz matrix (estimated): probability of answering quizzes correctly.</p>



Procedure

There were two requirements for the participants: 1) They need to answer the questionnaire in the role of a user to solve math quizzes in the system, and 2) They need to understand English. Finally, we recruited 18 participants who hold bachelor's degrees and conducted research in English. They consented to provide anonymized answers for research use.

Table 7 shows their background knowledge of high school math and expertise in recommender systems. As the mean values indicate, we consider they had sufficient math knowledge to make judgements in the role of a system user. Besides, they demonstrated different levels of expertise in recommender systems, which indicates they did not necessarily understand the algorithms behind recommender systems. Table 8 shows the orders of models these participants were presented with. The answers of P3, P5, P8, P15, P17 and P18 were excluded from the analysis as they only viewed the same explanations (only the first one or the first two) for all the three models, which were considered invalid evaluations. As a result, we used the answers from 12 participants for analysis and discussion.

Results

Table 9 shows how many of the explanations were viewed by the participants. Overall, participants viewed more explanations in CE+MF than in MF and Naïve CE. As the explanations of CE+MF combined the explanations of Naïve CE (Explanations 3-6) and MF (Explanations 7-10) in a linear order, we discuss the results by comparing [Naïve CE, MF], and [Naïve CE, CE+MF], respectively.

Table 7 Background information of the participants

Variable [scale]	N	Mean	SD	95% CI
Confidence in solving the recommended quiz [1-5]	12	4.333	0.985	[3.708, 4.959]
Familiarity with the terminologies related to recommender systems [1-5]				
Familiarity with "recommender system" [1-5]	12	4.417	0.669	[3.992, 4.841]
Familiarity with "collaborative filtering" [1-5]	12	3.500	1.446	[2.581, 4.419]
Familiarity with "matrix factorization" [1-5]	12	3.083	1.379	[2.207, 3.959]

Table 8 The order of evaluation

Order of evaluations	Participants
Naïve CE -> MF -> CE+MF	P1, P11, P13, P19
Naïve CE -> CE+MF -> MF	P2, P7, P14
MF -> CE+MF -> Naïve CE	P3, P8, P15
MF -> Naïve CE -> CE+MF	P4, P9
CE+MF -> Naïve CE -> MF	P5, P10, P17
CE+MF -> MF -> Naïve CE	P6, P12, P18

Table 9 Number of explanations viewed for the three models

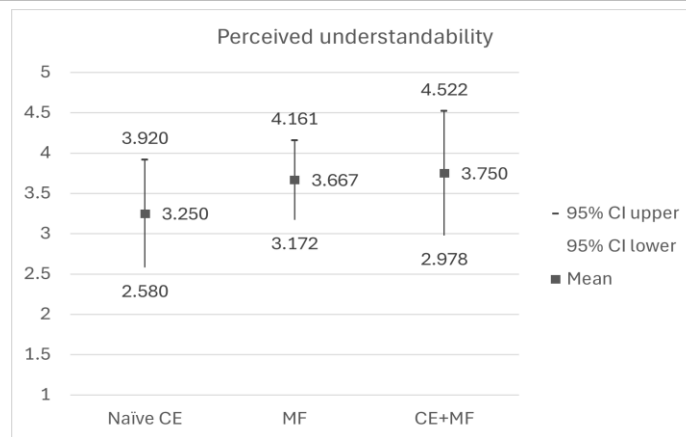
Participant	Naïve CE	MF	CE+MF
P1	4	5	5
P2	1	5	2
P4	3	5	5
P6	2	1	3
P7	5	5	10
P9	5	3	9
P10	1	3	1
P11	5	3	8
P12	7	5	4
P13	4	4	5
P14	4	5	10
P19	1	2	3
Average number of explanations viewed	3.500	3.833	5.417
(total number of explanations)	(7)	(5)	(10)

(1) Comparing Naïve CE and MF.

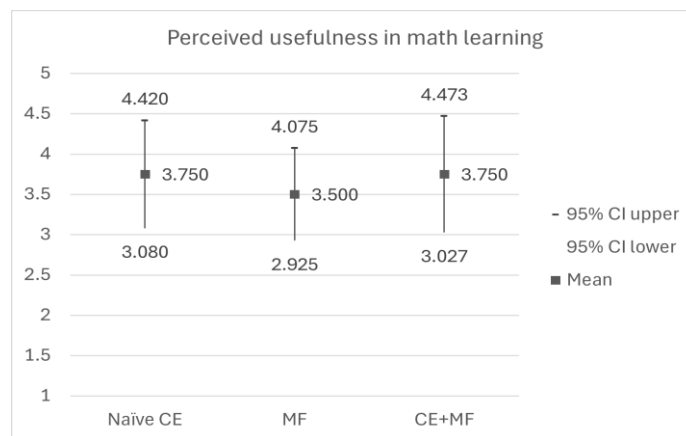
In Naïve CE, we found more participants stopped viewing explanations from the 4th and 5th explanations. As Explanations 1-4 do not include any computational details in the model, some of the participants may stop here out of satisfaction. Explanation 5 (“Your level of understanding of inequality is estimated from your attempts on other quizzes which also requires the knowledge of inequality. Specifically, you have got Quiz 120 wrong once (correctness rate=0/1=0.0), and you attempted Quiz 121 twice and got it right for the second time (correctness rate=1/2=0.5). Then, I estimate your level of understanding of inequality by taking the weighted average of the correctness rates.”) started to include computational details and demonstration with answering history. This could be overwhelming for the participants and stop them from further reading the explanations. In MF, more participants viewed through all the explanations. The reason that these participants did not stop at Explanations 1-4 could be polarized: Explanations 1-4 are confusing because they need more explanations to clarify, or the explanations are easy to follow. For either reason, the explanations in Naïve CE provide stop points for the participants while the ones in MF have a continuous flow to attract readers. However, it remains questionable whether the participants stopped due to positive reasons such as being convinced by the information or to a negative one such as losing interest. Similarly, it is interesting to explore whether the participants continued due to curiosity or confusion.

(2) Comparing Naïve CE and CE+MF.

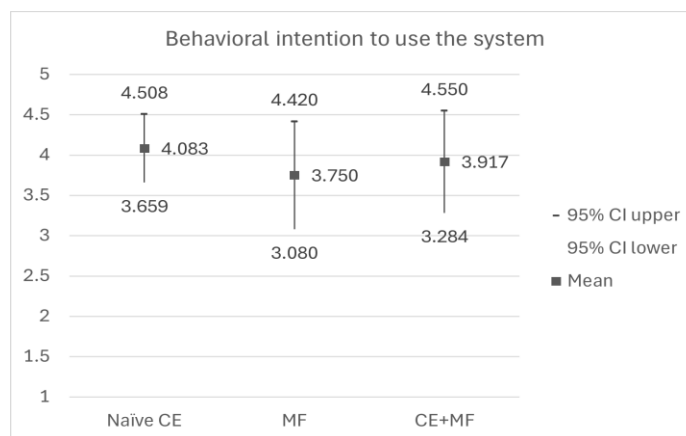
Participants viewed more explanations in CE+MF. It is interesting that both Naïve CE and CE+MF contained the same first 6 explanations but CE+MF was viewed deeper by the participants. One possible reason is that once the participant viewed Naïve CE part of explanations and entered MF part, they tend to read more explanations (see P7, P9 and P14).



(a) Statistics of perceived understandability of three models.



(b) Statistics of perceived usefulness in math learning of three models.



(c) Statistics of behavioral intention to use the system of three models.

Fig. 5 Evaluations of the explanations for the three models

Figure 5 shows the participants' evaluations on the explanations of three aspects. Interestingly, the evaluations demonstrated different trends of three models in perceived understandability, usefulness, and behavioral intention. Specifically, Naïve CE was the least, MF the second, and CE+MF the most understandable model (Figure 5a). Besides, Naïve CE and CE+MF were more useful in math learning and more motivating than MF (Figures 5b and 5c).

Naïve CE had the lowest score in perceived understandability but the highest scores in perceived usefulness and behavioral intention. In contrast, MF had a higher score in perceived understandability but lowest scores in perceived usefulness and behavioral intention. The participants did not necessarily fully understand how the system works but found the explanations useful in learning math, thus, were willing to use the system in a real setting. Or reversely, the participants somehow understood the mechanism of the system but did not find it very useful in learning math, thus, were less willing to use it.

CE+MF had the best overall evaluation of three aspects. CE+MF was more understandable than Naïve CE, equally useful in math learning, but not necessarily motivating for the participants to use in the real setting. This implies the intricate relationships among perceptions—understanding the recommendation and considering it useful in learning may not guarantee positive behaviors.

Discussion

RQ1: Is there a trade-off between the accuracy and the explainability?

As the results show, MF had the highest performance of estimating quiz mastery level while Naïve CE had the lowest, CE+MF had the medium performance. At the same time, Naïve CE and MF demonstrated different trends on different aspects of the explainability. Specifically, Naïve CE had lower perceived understandability and higher perceived usefulness in math learning and higher behavioral intention, while MF had the opposite results. Overall, CE+MF demonstrated high levels of explainability. Given the specific context and dataset, we did observe a trade-off between the accuracy and the explainability of recommender models, with the explainability being more intricate than we expected.

Why were the explanations of Naïve CE more difficult to understand than the ones of MF? One possible reason is the weakness of the mechanism of inherently explainable models. As found in Bell et al.'s (2022) study, an inherently explainable model—decision tree did not help the users to identify important features due to the bias brought by the tree diagram. In our survey, more participants stopped at Explanation 5 in Naïve CE model, which was lengthy and involved the estimation of mastery level of math concepts from learning history of other quizzes. The mechanism is relatively simple, and shallow compared with other black-box models, but not absolutely easy-to-understand for an end

user. In contrast, the explanations of MF encapsulated many complex details about mathematical reasoning, such as what the latent factors and matrix factorization are. Consequently, the explanations of MF were easier to “process” as they were simplified. As indicated in previous research, the detailedness of explanations may affect its effects (Chatti et al., 2022; Kulesza et al., 2013). Chatti et al. (2022) found that the perception of explanations with different levels of detailedness was affected by the explanation goal and user type. In this context, the participants were supposed to improve their math learning by solving quizzes in the system. In this sense, their primary goal is not to understand the mathematical background of the system with a large cognitive cost. Kulesza et al. (2013) proposed a framework to model the soundness and completeness of the explainability, which may be a future direction to refine the explanations.

Why did the participant show positive judgements towards Naïve CE without fully understanding how it worked? As discussed previously, the primary goal of the user in this context is to improve learning by solving recommended quizzes. Understanding the model’s mechanism is not equal to obtaining the necessary information for the task. One of the participants mentioned in the open question that s/he would like to know “what if i still cannot solve the recommended problem, because I failed all the problems on these 2 knowledge. there can be also high possibility to cannot solve the recommended problem with wheel spinning.”. Obviously, the participant still had questions about how the system works, but the information presented in the explanations helped her/him to consider a learning scenario. The informativeness of explanations may be related to the development of meta-cognitive skills (Dai et al., 2024), which needs to be further explored. In contrast, two participants mentioned they wanted to know what the “factors” of MF model are, so that they can make a judgement on the selection of quizzes. In the real world, it is more important to provide pedagogically useful information than to merely explain how the model works.

RQ2: Is it feasible to enhance the accuracy of inherently explainable models by integrating them with black-box models?

As we did observe a trade-off between the accuracy and the explainability of three models, it leads to the question of how to develop an explainable and accurate recommender model. Our experiment served as an example to improve the accuracy of an inherently explainable model. The advantages of this approach are as follows: 1) The advantages of explanations in inherently explainable models are difficult to generate from black-box models. For example, the math concepts used in Naïve CE played an important role in providing math-related information, which are not the original idea of MF. 2) The integrated model preserved both characteristics of two models. For example, CE+MF built on the basic framework of Naïve CE and integrated MF into a local step in its mechanism. As a result,

the out-layer explanations of CE+MF provided math-related information, and the accuracy improved. Theoretically, a model that has a balanced performance of accuracy and explainability should be preferred (Khosravi et al., 2022; Molnar et al., 2022). However, it remains challenging to quantify the balance of explainability and accuracy. This is closely dependent on the context, in this case, improving learning effects. In some extreme cases, it would be sufficiently explainable if the students are satisfied with the general difficulty without the interest in understanding how it is computed. Therefore, it is necessary to select models based on the purpose of the model and explanations.

Conclusion and future work

In this study, we explored the accuracy and the explainability of math recommender systems. Focusing on three recommender models—an inherently explainable model (Naïve CE), a black-box model (MF), and an integrated model (CE+MF), the accuracy was evaluated by measuring how correctly the model can estimate students' mastery level of the quizzes; the explainability was evaluated from three perspectives in a questionnaire survey. The findings indicate a trade-off between the accuracy and explainability of the recommender models. However, the explainability was more complex and dependent on the context. In the learning scenario, participants found the model useful in math learning without fully understanding how the model worked. Overall, the integrated model displayed a balanced level of accuracy and explainability, which implies the feasibility to develop an explainable educational recommender system by improving the accuracy of an inherently explainable model.

Some limitations can be addressed in future work: 1) In this study, we measure accuracy at the step of estimating students' probabilities to succeed in answering quizzes. The process to generate recommendations based on the estimation was not included yet, as the "goodness" of a recommendation is more difficult to measure. Future work should be conducted to extend the measurement of accuracy of such educational intelligent systems. 2) The sample size was limited to provide statically strong evidence in the questionnaire survey. In the future work, we plan to conduct online experiments with students in authentic environments. 3) In this study, we only explored one way to integrate Naïve CE and MF. There exist other ways to improve the accuracy of Naïve CE by elaborating the domain model, or to improve the explainability of MF by feeding explicit factors. It is promising to improve accuracy and explainability at the same time. 4) The quantification of practical effects of accuracy and explainability is still unclear. When we develop explainable educational intelligent systems, how can we compare the learning effects between an improvement of accuracy and an improvement of explainability? We will explore ways to address this issue.

Abbreviations

AI: Artificial Intelligence; Naïve CE: Naïve Concept Explicit; MF: Matrix Factorization; CE+MF: Concept-Explicit Matrix Factorization; XAI: Explainable Artificial Intelligence; AUC: Area Under ROC Curve; RMSE: Root Mean Square Error.

Acknowledgements

We would like to thank the participants in the survey for their time in helping to evaluate the performance of the method proposed in this research. Without their time and cooperation, this research would not be possible.

Authors' contributions

YD implemented the system, designed the experiment, performed data analysis and drafted the initial manuscript. BF provided support for system implementation, experiment design, data analysis, and edited the manuscript. HO was responsible for funding acquisition and supervision. All authors read and approved the final manuscript.

Authors' information

Yiling Dai is a Program-Specific Researcher at the Academic Center for Computing and Media Studies, Kyoto University. She received a bachelor's degree from Zhejiang University, a master's degree from the Graduate School of Business, Rikkyo University, and a PhD degree from the Graduate School of Informatics, Kyoto University. Her research interests include: Information Retrieval, Knowledge Discovery, Educational Data Mining and Learning Analytics.

Brendan Flanagan is an Associate Professor at the Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, and the Graduate School of Informatics at Kyoto University. He received a bachelor's degree from RMIT University and master's and Ph.D. degrees from the Graduate School of Information Science and Electrical Engineering, Kyushu University. His research interests include: Learning Analytics, Educational Data Science, Educational Data Mining, NLP/Text Mining, Machine Learning, Computer Assisted Language Learning, and the Application of Blockchain in Education.

Hiroaki Ogata is a full Professor at the Academic Center for Computing and Media Studies, Kyoto University. His research interests include: Learning Analytics, Evidence-Based Education, Educational Data Mining, Educational Data Science, Computer Supported Ubiquitous and Mobile Learning, and CSCL.

Funding

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, 23H01001, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, NEDO JPNP20006.

Availability of data and materials

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan

² Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University, Kyoto, Japan

Received: 28 February 2024 Accepted: 4 November 2024

Published online: 1 January 2026 (Online First: 3 March 2025)

References

- Abdi, S., Khosravi, H., & Sadiq, S. (2018). Predicting student performance: The case of combining knowledge tracing and collaborative filtering. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the International Conference on Educational Data Mining* (pp. 545–548). International Educational Data Mining Society.
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Alonso, J. M., & Casalino, G. (2019). Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In D. Burgos, M. Cimitile, P. Ducange, R. Pecori, P. Picerno, P. Raviolo & C. M. Stracke (Eds.), *Higher education learning methodologies and technologies online* (pp. 125–138). Springer International Publishing. https://doi.org/10.1007/978-3-030-31284-8_10
- Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Benoitot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies,

- opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Barnes, T. (2005). The Q-matrix method: Mining student response data for knowledge. In *Proceedings of American Association for Artificial Intelligence 2005 Educational Data Mining Workshop* (pp. 39–46). Association for the Advancement of Artificial Intelligence. <http://www.aaai.org/Papers/Workshops/2005/WS-05-02/WS05-02-006.pdf>
- Barria-Pineda, J., Akhuseyinoglu, K., Želem-Čelap, S., Brusilovsky, P., Milicevic, A. K., & Ivanovic, M. (2021). Explainable recommendations in a personalized programming practice system. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin & V. Dimitrova (Eds.), *Artificial Intelligence in Education. AIED 2021. Lecture Notes in Computer Science*, vol 12748 (pp. 64–76). Springer, Cham. https://doi.org/10.1007/978-3-030-78292-4_6
- Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 248–266). ACM. <https://doi.org/10.1145/3531146.3533090>
- Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education*, 24(5), 442–459.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
- Chatti, M. A., Guesmi, M., Vorgerd, L., Ngo, T., Joarder, S., Ain, Q. U., & Muslim, A. (2022). Is more always better? The effects of personal characteristics and level of detail on the perception of explanations in a recommender system. In A. Bellogin, L. Boratto, O. C. Santos, L. Ardisson & B. Knijnenburg (Eds.), *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 254–264). ACM. <https://doi.org/10.1145/3503252.3531304>
- Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298, 103503. <https://doi.org/10.1016/j.artint.2021.103503>
- Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278. <https://doi.org/10.1007/BF01099821>
- Dai, Y., Flanagan, B., Takami, K., & Ogata, H. (2022). Design of a user-interpretable math quiz recommender system for Japanese high school students. In B. Flanagan, R. Majumdar, H. Li, A. Shimada, F. Okubo & H. Ogata (Eds.), *Proceedings of the 4th Workshop on Predicting Performance Based on the Analysis of Reading Behavior - DC in LAK22 Co-Located with 12th International Learning Analytics and Knowledge Conference (LAK22)*, 3120 (pp. 30–38). CEUR-WS.org. <https://ceur-ws.org/Vol-3120/>
- Dai, Y., Takami, K., Flanagan, B., & Ogata, H. (2022). Investigation on practical effects of the explanation in a K-12 math recommender system. In S. Iyer et al. (Eds.), *Proceedings of the 30th International Conference on Computers in Education* (pp. 7–12). Asia-Pacific Society for Computers in Education.
- Dai, Y., Takami, K., Flanagan, B., & Ogata, H. (2024). Beyond recommendation acceptance: Explanation's learning effects in a math recommender system. *Research and Practice in Technology Enhanced Learning*, 19, 020. <https://doi.org/10.58459/rptel.2024.19020>
- Desmarais, M. C., & Pelczar, I. (2010). On the faithfulness of simulated student performance data. In R. S. J. de Baker, A. Merceron & P. I. P. Jr (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 21–30). International Educational Data Mining Society. https://educationaldatamining.org/EDM2010/uploads/proc/edm2010_submission_56.pdf
- Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education*, 106, 83–96. <https://doi.org/10.1016/j.compedu.2016.12.001>
- Flanagan, B., Takami, K., Takii, K., Dai, Y., Majumdar, R., & Ogata, H. (2021). EXAIT: A symbiotic explanation learning system. In M. M. T. Rodrigo et al. (Eds.), *Proceedings of the 29th International Conference on Computers in Education* (pp. 404–409). Asia-Pacific Society for Computers in Education.
- Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 12(3), 31–54. <https://doi.org/10.5281/zenodo.4143614>
- Guleria, P., & Sood, M. (2023). Explainable AI and machine learning: Performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Education and Information Technologies*, 28(1), 1081–1116. <https://doi.org/10.1007/s10639-022-11221-2>
- Hur, P., Lee, H., Bhat, S., & Bosch, N. (2022). Using machine learning explainability methods to personalize interventions for students. In A. Mitrovic, N. Bosch, A. I. Cristea & C. Brown (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 438–445). International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.6853181>
- Khosravi, H., Cooper, K., & Kitto, K. (2017). RiPLE: Recommendation in Peer-Learning Environments Based on Knowledge Gaps and Interests. *Journal of Educational Data Mining*, 9(1), 42–67. <https://doi.org/10.5281/zenodo.3554627>
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, 3, 100074. <https://doi.org/10.1016/j.caeai.2022.100074>
- Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30–37. <https://doi.org/10.1109/MC.2009.263>

- Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In O. Brdiczka, P. Chau, G. Carenini, S. Pan & P. O. Kristensson (Eds.), *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 126–137). ACM. <https://doi.org/10.1145/2678025.2701399>
- Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing* (pp. 3–10). IEEE. <https://doi.org/10.1109/VLHCC.2013.6645235>
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 464–470. <https://doi.org/10.1016/j.tics.2006.08.004>
- Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9), 1315–1316.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Molnar, C., Köhig, G., Herbringer, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2022). General pitfalls of model-agnostic interpretation methods for machine learning models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 39–68). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_4
- Ooge, J., Kato, S., & Verbert, K. (2022). Explaining recommendations in e-learning: Effects on adolescents' trust. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (pp. 93–105). ACM. <https://doi.org/10.1145/3490099.3511140>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Swamy, V., Radmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. In A. Mitrovic, N. Bosch, A. I. Cristea & C. Brown (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 98–109). International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.6852964>
- Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2008). Matrix factorization and neighbor based algorithms for the netflix prize problem. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08)* (pp. 267–274). ACM. <https://doi.org/10.1145/1454008.145404>
- Takami, K., Dai, Y., Flanagan, B., & Ogata, H. (2022). Educational explainable recommender usage and its effectiveness in high school summer vacation assignment. In A. F. Wise, R. Martinez-Maldonado & I. Hilliger (Eds.), *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK22)* (pp. 458–464). ACM. <https://doi.org/10.1145/3506860.3506882>
- Takami, K., Flanagan, B., Dai, Y., & Ogata, H. (2023). Toward trustworthy explainable recommendation: Personality based tailored explanation for improving e-learning engagements and motivation to learn. In I. Hilliger, H. Khosravi, B. Rienties & S. Dawson (Eds.), *The Companion Proceedings of the 13th International Conference on Learning Analytics & Knowledge* (pp. 120–122). The Society for Learning Analytics Research.
- Tintarev, N., & Masthoff, J. (2007). A survey of explanations in recommender systems. In *Proceedings of 2007 IEEE 23rd International Conference on Data Engineering Workshop* (pp. 801–810). IEEE. <https://doi.org/10.1109/ICDEW.2007.4401070>
- Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- Vultureanu-Albiş, A., & Bădică, C. (2022). A survey on effects of adding explanations to recommender systems. *Concurrency and Computation: Practice and Experience*, 34(20), e6834. <https://doi.org/10.1002/cpe.6834>
- Wang, N., Wang, H., Jia, Y., & Yin, Y. (2018). Explainable recommendation via multi-task learning in opinionated text data. In K. Collins-Thompson, Q. Mei, B. Davison, Y. Liu & E. Yilmaz (Eds.), *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)* (pp. 165–174). ACM. <https://doi.org/10.1145/3209978.3210010>
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. *Educational Measurement*, 4, 111–153.
- Yeung, A. A. (2020). Matrix-factorization-in-python. In *GitHub repository*. GitHub. <https://github.com/albertayueng/matrix-factorization-in-python>
- Yu, R., Pardos, Z., Chau, H., & Brusilovsky, P. (2021). Orienting students to course recommendations using three types of explanation. In J. Masthoff, E. Herder, N. Tintarev & M. Tkalčič (Eds.), *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 238–245). ACM. <https://doi.org/10.1145/3450614.3464483>
- Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1), 1–101. <https://doi.org/10.1561/1500000006>

Publisher's Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research and Practice in Technology Enhanced Learning (RPTEL)
is an open-access journal and free of publication fee.