User-centric Evaluation of GenAI Alignment and Recommendations based on Predictive Learning Analytics*

Hesham Ahmed[1], Halil Kayaduman[1,2], Sonsoles López-Pernas[1], Markku Tukiainen[1], and Mohammed Saqr[1]

[1] *University of Eastern Finland, Yliopistokatu 2, 80100 Joensuu, Finland*
[2] *Distance Education Application and Research Center, Inonu University, Malatya, Turkey*

Abstract

Predictive models are one of the hallmarks of learning analytics research, relying on learner data to predict academic achievement and dropouts, enabling targeted interventions. Using a user-centric evaluation framework, we assessed the recommendations generated by ChatGPT based on the results of 136 studies that used student data for predictive modeling. The evaluation considered general attributes (accuracy, coherence, justification) as well as education-specific criteria (alignment with learning theories, ethics, learner-centeredness). The results indicate that, while LLM-generated recommendations are generally accurate, coherent and useful, they often lack alignment with diverse learning theories and fail to address inclusivity and higher-order cognitive skills effectively. Therefore, to operationalize LLMs to provide automated feedback to students, these aspects should be explicitly considered in the prompt design.

Keywords

Generative AI, Predictive Learning Analytics, Large Language Models, User-centric Evaluation

1. Introduction

Generative Artificial Intelligence describes a set of computational techniques that can generate mostly comprehensible content in the form of text, image, and video that is new out of training data [1]. Large Language Models (LLMs) are a subset of such techniques that is based on neural networks which is trained on hundreds of terabytes of textual data [2]. An example of such models is ChatGPT, which has demonstrated human-like performance on a wide range of natural-language oriented objectives ranging from translation, writing intelligible essays, and creating functional code [3]. This, in return, has encouraged many to explore its possible benefits in the realm of education [4].

Some studies have explored the utility of LLMs-powered Recommendation Systems in educational-related contexts [5], [6]. While it is early as the adoption is ramping up, it is necessary to understand their impact not only from a functional standpoint [7], but also in terms of user experience and alignment with pedagogical objectives. Teachers, students, and other stakeholders will use these tools in varied ways, underlining the need for user-centric evaluation to ensure that the recommendations generated are of a high quality from different aspects like implementability, alignment with learning theories, and ethicality.

Such recommendation systems can be built to support certain objectives, for example, [5] aimed to support student learning recommendations using, among others, Knowledge Graph Contextualization. In our case, the recommendations are based on predictive Learning Analytics (LA) models [8]. Predictive LA is concerned with using learner-related data to create predictions of possible future scenarios to aid in making interventions that avoid the negative ones. Such

Second International Workshop on Generative AI for Learning Analytics, 2025.
hesham.ahmed@uef.fi (H. Ahmed); halilkayaduman@gmail.com (H. Kayaduman); sonsoles.lopez@uef.fi (S. López-Pernas); markku.tukiainen@uef.fi (M. Tukiainen); mohammed.saqr@uef.fi (M. Saqr)
0009-0005-7042-4938 (H. Ahmed); 0000-0001-5316-1893 (H. Kayaduman); 0000-0002-9621-1392 (S. López-Pernas); 0000-0002-8630-5248 (M. Tukiainen); 0000-0001-5881-3109 (M. Saqr)

predictions are a result of revealing statistical correlative relations between different features like previous academic performance, current credit load, and behavior on learning management systems.

Evaluating LLMs has generally been through rigorous frameworks [9] that have focused on evaluating its performance automatically using standardized datasets and benchmarks. In a study similar to ours, [10] have evaluated LLM-powered recommendation systems using both objective measures and user-centric subjective criteria based on a revised version of the ResQue framework [11]. However, the recommendation system was concerned with leisure-related events and activities and did not relate to education. Recently, we have evaluated recommendations from a Retrieval-Augmented Generation (RAG) system. RAG is a set of methods that enhances the quality of the LLMs responses through supplying it with additional external knowledge [12]. This RAG relied on predictive LA models extracted from state-of-the-art research on LA. The responses were generally more specific compared to the responses of a typical LLM. However, the responses were not very accurate in many cases and lacked precision.

In this study, we aim to comprehensively evaluate the quality of recommendations provided by LLMs based on their interpretation of learning analytics research findings. This is because the ability to offer recommendations rests on the ability to digest and translate research findings into actual practical recommendations that account for different criteria as in some instances, the resulting recommendations could be not only unintelligible but also potentially harmful.

2. Methodology

In this study, we aim to follow the steps shown in Figure 1 to answer the following questions:

1. How accurate are LLMs in interpreting predictive learning analytics results and providing useful recommendations to students?

2. How aligned are the recommendations with learning theories, learner-centeredness, ethics and engagement with higher order cognitive skills?

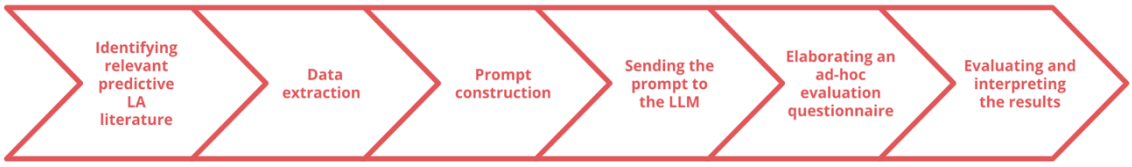

Figure 1: The methodology followed in the study

2.1. Studies collection

The first step of this process was to collect all predictive LA research through snowballing from existing systematic literature reviews that used student data to create predictive models on student achievement, retention, success and all other students' outcomes. First, we identified a total of 13 relevant systematic literature reviews (see Figure 2). The second step was mining the references of the systematic reviews and compiling them into a list of 1,517 references. After eliminating duplicate entries, non-English articles, and those published before 2011, by examining the title, abstract, and keywords, a total of 476 articles remained. Next, each article was manually inspected to verify whether they used student data (such as learning management system activity) to predict a target (such as grades) and reported results that displayed the correlation between the predictors/features and the predictor in an interpretable format (such as piece of text, a table, or a figure). To account for possible mistakes and misses, the inclusion and exclusion procedure was validated by a second researcher. At the end, a total of 136 articles fulfilled all the criteria and were passed for later stages. The flow of this process is illustrated in Figure 2.

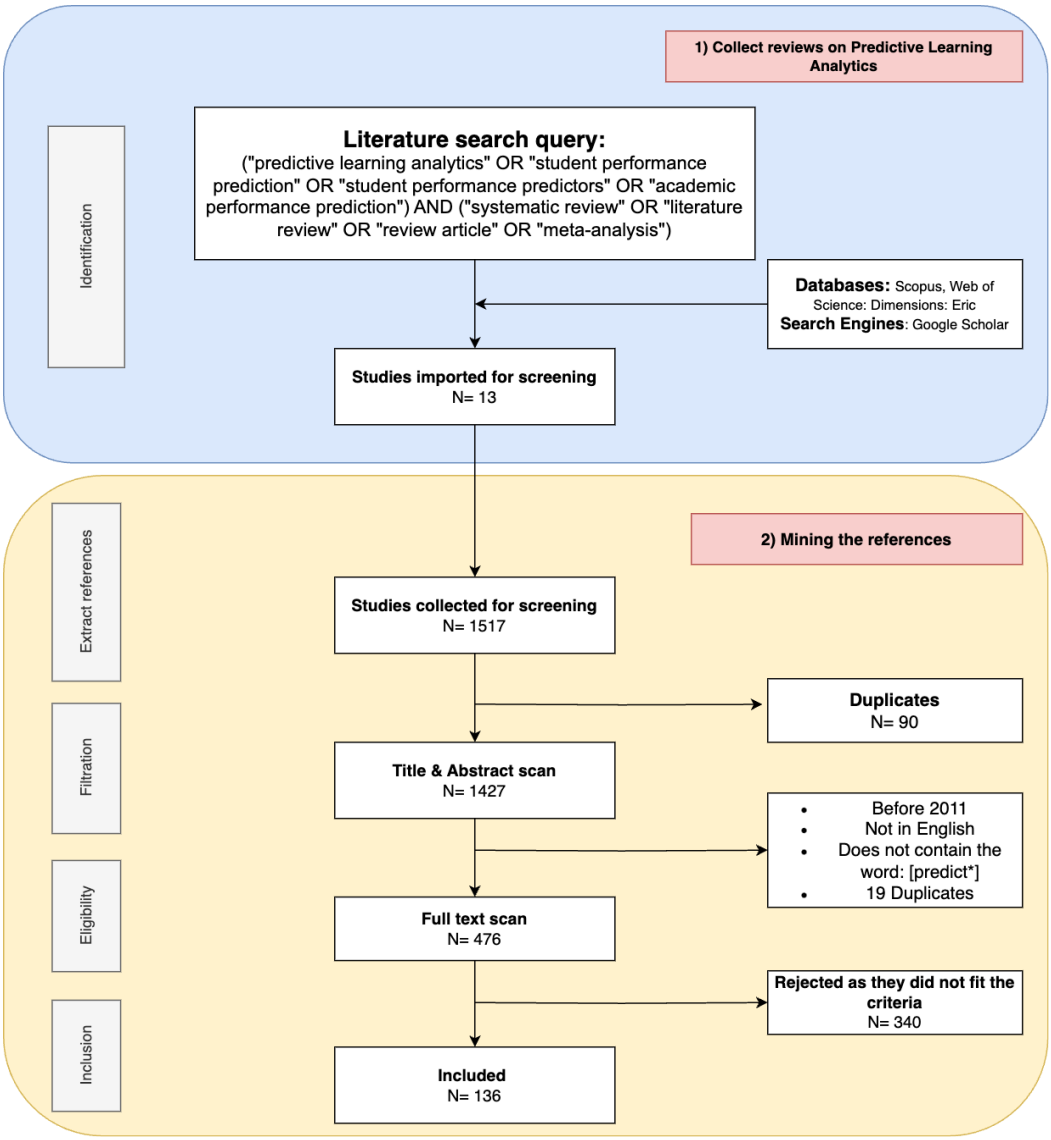

Figure 2: The studies articles collection process

2.2. Data extraction

The following data were collected manually out of each study in a tabular format: study title, year of publication of the study, level of education, type of study (STEM/ Non-STEM), duration of the data collection, number of students, data sources (features), description of the features (if available), targeted variable, prediction method, and format of the results (table/graph/text). The targeted variable, the features and their description, and the statistical model that describes the relation between the predictors and the predicted were also collected as a screenshot in JPEG and PNG file formats. The rationale for this is that the presentation formats such data was described in each study were rarely the same. So, this will aid in standardizing the format of the input to the LLM. Such screenshots were taken in the highest possible resolution to avoid misinterpretation and hallucination.

The dominant **level of education** targeted in the study was university-level representing approximately 74%. STEM represented the largest portion of **types of study** with 55.1% followed by mixed types with 28.7%. The median **duration of the data collection** was around 1 year. The **number of students** in the studies was mostly below 1000. Figure 3 shows the years in which the studies collected were published.

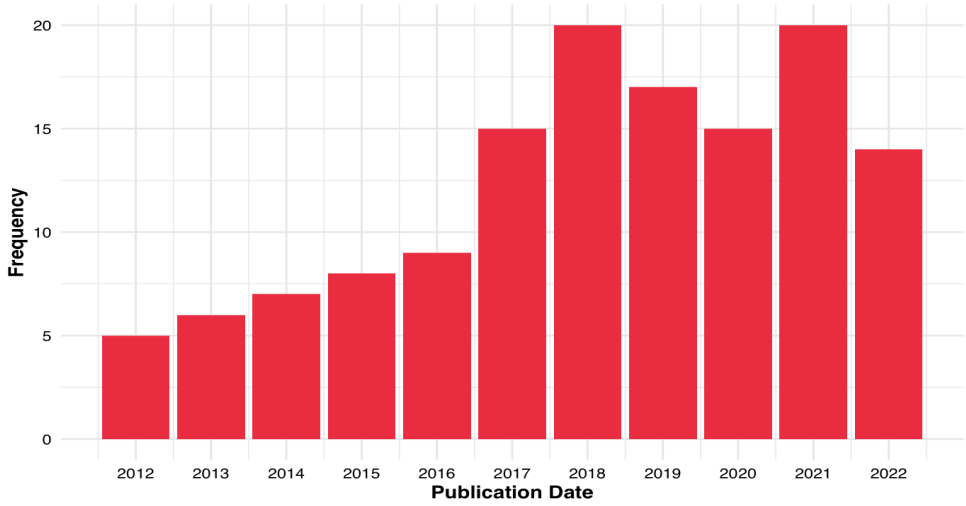

Figure 3: This figure shows the frequency of studies collected by year of publication

2.3. LLM prompting and response collection

A prompt was created to cover each of the 136 studies. We followed the Basic Prompting technique [13], that is, supplying the LLM with the input and a request without giving it examples of the expected output as a guide. The prompt was meant to be general without suggestive language as we wanted not to prime its output towards a specific shape or form or anchor it to consider any specific criteria. ChatGPT4 was used to create the recommendations not only because it is one of the most advanced and commonly used but also it allowed multiple images as an input. After creating the prompts, each prompt was supplied to ChatGPT4 with each in a separate conversation to avoid contextual overlap. Afterwards, the responses were collected as text to facilitate its evaluation. The format of the prompt is shown in Table 1.

Table 1
Description of the different parts of the standardized prompt

Section	Description	Format
Generic Instructions	This part describes the objectives of the LLM. It supplies it with a context and a request and some clarifying instructions. This is its fixed form: *"Your objective: I have conducted research and based on the findings, could you interpret the results and provide practical, actionable recommendations or strategies for students? Please, explain the reasoning behind each single recommendation. In your text, replace the abbreviations with their full form if applicable. Please, use the table of abbreviations if provided and when needed. Do not add any additional text, only recommendations."*	Text
Research Title	This part shows the title of the study.	Text
Research Aim	This part describes the objective from the research.	Text
Research Questions	This part shows the research questions of the study, only if explicitly mentioned.	Text
Data Analysis	This part describes the data and how it was analyzed to create a predictive model.	Text
Screenshots/data source	These are screenshots from the study that describes the predictors/features and its relation to the predicted in a predictive model. The data visualization can be in different formats: Bar chart, table, correlation matrix, etc.	PNG / JPEG

2.4. Evaluation framework development

To evaluate the responses, a questionnaire, with the name of *Learning Analytics Recommendations Alignment Questionnaire* (LARAQ), was developed that relied on criteria that address general attributes of recommendation systems and criteria that address learner-related attributes. For the general criteria, the ResQue framework has been widely used to allow the users of a recommendation system to subjectively assess it holistically [10]. This framework has been utilized to assess recommendation systems in different domains such as music [14] and movies [15]. Similarly, [16] has cultivated a framework to evaluate conversational recommendation systems specifically. The general criteria that were chosen are: Recommendation Accuracy, Subjective measure of the presentability of the recommendation, Justification, Perceived Usefulness, Consistency & Coherence.

Learner-related attributes assess that the recommendations are not only applicable but also how much it adheres to pedagogical and ethical principles. The criteria are: Implementability & practicality, Privacy & ethicality, Alignment with learning Theories, Diversity, Equity, Inclusion, Learner-Centeredness, Engagement with Higher-Order Cognitive Skills. The criteria in LARAQ were extracted from different frameworks and evaluations and adapted to the context of education. Two evaluators independently assessed the evaluation. For quantitative questions, the average score was calculated, whereas qualitative questions were evaluated through consensus. Tables 2 and 3 show each criterion and their respective questions alongside its sources.

Table 2
This table shows General criteria of LARAQ with their respective questions, scale, and references.

Target	Question	Scale	References
General criteria			
1. Recommendation Accuracy	Do the generated recommendations match with the provided paper?	5-point Likert	[11], [10], [17], [18], [16], [19], [14], [20], [21], [22]
2. Subjective measure of the presentability of the recommendation	Is the structure of the recommendations clear and well-organized?	5-point Likert	[11], [16]
3. Justification	Are the reasons for the suggestions clearly explained?	5-point Likert	[11], [16], [20]
4. Perceived Usefulness	How beneficial are the recommendations?	5-point Likert	[11], [10], [16], [19], [23]
5. Consistency & coherence	Do the suggestions maintain a coherent and consistent line of thought?	5-point Likert	[10]

Table 3
This table shows Learner-related criteria of LARAQ with their respective questions, scale, and references.

Target	Question	Scale	References
Learner-related criteria			
6. Implementability & practicality	How feasible are the recommendations to put into practice?	5-point Likert	[11], [24], [25]
7. Privacy & ethicality	Do the recommendations suggest using protected data?	Yes/No	[11], [24], [26], [27]
8. Alignment with learning Theories	Are the recommendations grounded in educational theories?	Multiple choice	[28], [29], [24], [30]
9. Diversity	To what extent do the recommendations consider diverse student backgrounds and perspectives?	5-point Likert	[31]
10. Equity	Do the recommendations address possible disparities for disadvantaged groups?	5-point Likert	[32], [31]
11. Inclusion	To what degree do the recommendations foster an inclusive learning environment for all the students?	5-point Likert	[33], [31]
12. Learner-Centeredness	Do the recommendations prioritize the learners' needs?	5-point Likert	[24], [34]
13. Engagement with Higher-Order Cognitive Skills	Do the recommendations promote advanced cognitive skills?	5-point Likert	[29], [35]

3. Results

The box plot in Figure 4 shows the evaluations of Questions 1-6 and Questions 9-13. It reveals that Accuracy, Presentability, Justification, Usefulness, Consistency & Coherence, and Practicality receive higher median ratings (around 4.0 to 4.5), indicating generally positive assessments. On the other hand, Diversity, Equity, Inclusion, and Engagement with Higher-Order Cognitive Skills have lower median ratings (around 2.0 to 3.0). Lastly, Learner-Centeredness shows moderate ratings with some variability.

Figure 5 illustrates the frequency of the learning theories that the recommendations were aligned with. The X-Axis lists the learning theories, while the Y-Axis shows their frequency, ranging from 0 to around 90. The graph reveals notable variations in the prominence of these theories. Constructivism, Cognitivism, Behaviorism in order are the most frequently mentioned theories with scores of almost 90, around 75, and almost 60, respectively. With slightly above 40, Motivation theories follow the order. A mid-range cluster includes Social Learning, Humanism, Situated Learning, Metacognition, and Self-Regulated Learning, all hovering around 20. In contrast, several learning theories appear much less frequently, Connectivism and Transformative Learning register frequencies below 5 each. Overall, the data suggests that traditional theories like Behaviorism and Cognitivism dominate the landscape of learning recommendations, while emerging or specialized

theories are referenced far less often. Lastly, Question 8 showed an overwhelming majority of recommendations (97.1) did not suggest using protected information according to the GDPR while a very small minority (2.9%) did.

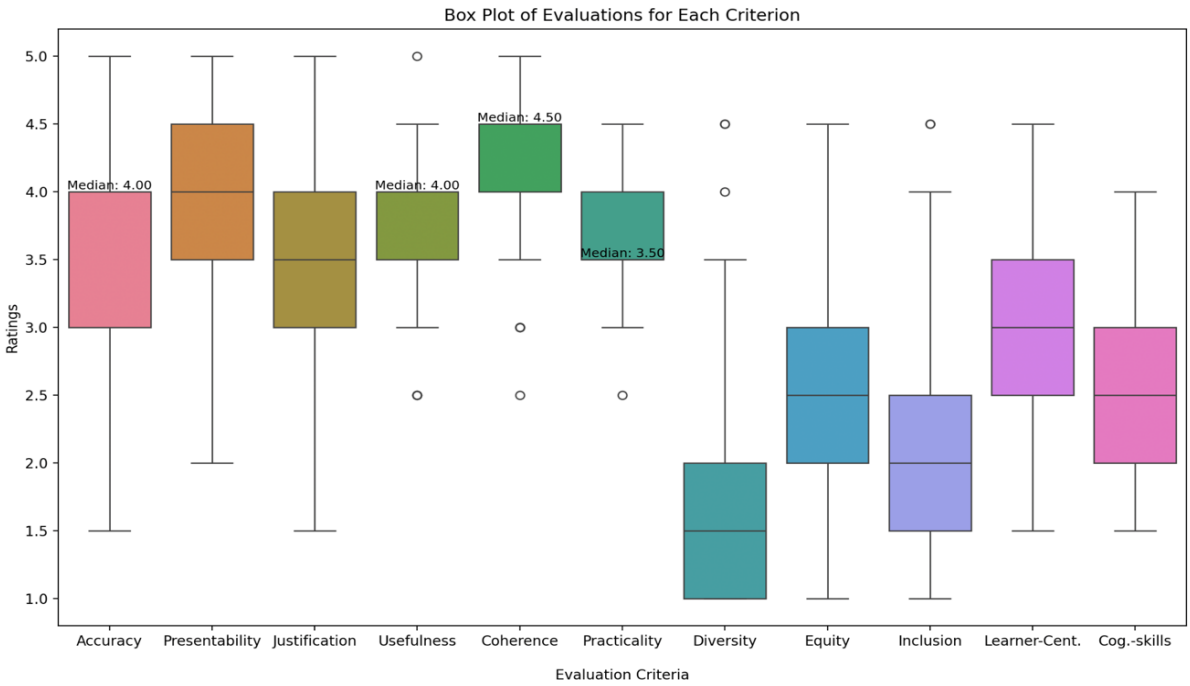

Figure 4: Box Plot illustration of Questions 1-6 and Questions 9-13

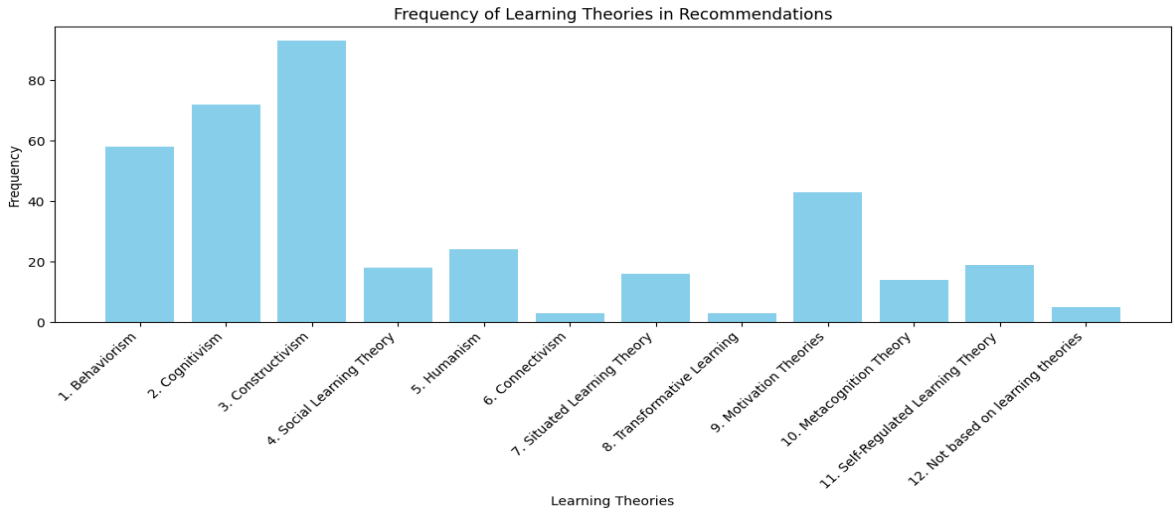

Figure 5: The frequency answers to Question 7: Alignment with learning theories

4. Discussion and conclusion

The recommendations were generally perceived as relevant to the supplied papers, succeeding in mentioning the most important feature. However, in cases where the number of features supplied was high, due to the LLM's output size limit, a lot of the features were neglected. The recommendations were well-structured and easy to read as well as beneficial and essential for improving outcomes. The clarity in reasoning was generally good as a rationale for the recommendations was mostly present. The suggestions were practical with some recommendations being vague and hard to implement. The arguments were understandable and the logic holding the recommendations together was consistent. A noticeable decline appears when examining learning-related criteria. These results suggest that the recommendations inadequately consider diverse

backgrounds and needs of different disadvantaged groups or foster inclusivity as they both hold the two lowest medians. Furthermore, promotion of higher-order cognitive processes such as synthesis and evaluation were insufficient. Additionally, the recommendations did not suggest using any sensitive data (according to the GDPR) of the learners if it was not included in the data of the supplied paper. Finally, the recommendations are well-grounded in learning theories.

The results suggest the prompt should be crafted with emphasis on the learner-related criteria by explicitly mentioning it while the LLM seems to perform well in understanding the tasks and in formatting the recommendation logically and aesthetically. Furthermore, in the absence of a description of the features, the LLM struggled to infer the meaning of some features from their names solely. Instead, it attempted to guess its meaning from the context and in many cases it either failed in its interpretation or took the safe route and did not include such ambiguous features in the recommendations.

For future work, we plan to evaluate LLMs fine-tuned with educational datasets. Moreover, we plan to use raw results and individual predictions for each student, combined with eXplainable AI methods that provide explanations for the predictions. This approach aims to offer personalized insights, addressing some of the gaps identified in our current study.

Acknowledgements

The paper is co-funded by the Academy of Finland (Suomen Akatemia) Research Council for the project "Towards precision education: Idiographic learning analytics (TOPEILA)", Decision Number 350560 which was received by the last author, and the project "Optimizing Clinical Reasoning in Time-Critical Scenarios: A data-driven multimodal approach (CRETIC)", Decision Number 360746, which was received by the third author.

Declaration on Generative AI

The author(s) have not leverage Generative AI tools when preparing the manuscript.

References

[1] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," *Bus. Inf. Syst. Eng.*, vol. 66, no. 1, pp. 111–126, Feb. 2024, doi: 10.1007/s12599-023-00834-7.

[2] M. Shanahan, "Talking about Large Language Models," *Commun. ACM*, vol. 67, no. 2, pp. 68–79, Feb. 2024, doi: 10.1145/3624724.

[3] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Educ. Sci.*, vol. 13, no. 4, p. 410, 2023.

[4] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.

[5] H. Abu-Rasheed, M. H. Abdulsalam, C. Weber, and M. Fathi, "Supporting Student Decisions on Learning Recommendations: An LLM-Based Chatbot with Knowledge Graph Contextualization for Conversational Explainability and Mentoring," Jan. 24, 2024, *arXiv*: arXiv:2401.08517. doi: 10.48550/arXiv.2401.08517.

[6] J. Liao *et al.*, "LLaRA: Large Language-Recommendation Assistant," May 04, 2024, *arXiv*: arXiv:2312.02445. doi: 10.48550/arXiv.2312.02445.

[7] D. D. Palma, G. M. Biancofiore, V. W. Anelli, F. Narducci, T. D. Noia, and E. D. Sciascio, "Evaluating ChatGPT as a Recommender System: A Rigorous Approach," Jun. 04, 2024, *arXiv*: arXiv:2309.03613. doi: 10.48550/arXiv.2309.03613.

[8] C. Brooks and C. Thompson, "Predictive modelling in teaching and learning," *Handb. Learn. Anal.*, pp. 61–68, 2017.

[9] Y. Chang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.

[10] H. Kunstmann, J. Ollier, J. Persson, and F. von Wangenheim, "EventChat: Implementation and user-centric evaluation of a large language model-driven conversational recommender system for exploring leisure events in an SME context," Jul. 09, 2024, *arXiv*: arXiv:2407.04472. doi: 10.48550/arXiv.2407.04472.

[11] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in *Proceedings of the fifth ACM conference on Recommender systems*, Chicago Illinois USA: ACM, Oct. 2011, pp. 157–164. doi: 10.1145/2043932.2043962.

[12] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar. 27, 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[13] J. Shin, C. Tang, T. Mohati, M. Nayebi, S. Wang, and H. Hemmati, "Prompt Engineering or Fine Tuning: An Empirical Assessment of Large Language Models in Automated Software Engineering Tasks," Oct. 11, 2023, *arXiv*: arXiv:2310.10508. doi: 10.48550/arXiv.2310.10508.

[14] Y. Jin, W. Cai, L. Chen, N. N. Htun, and K. Verbert, "MusicBot: Evaluating Critiquing-Based Music Recommenders with Conversational Interaction," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, Beijing China: ACM, Nov. 2019, pp. 951–960. doi: 10.1145/3357384.3357923.

[15] F. Pecune, S. Murali, V. Tsai, Y. Matsuyama, and J. Cassell, "A Model of Social Explanations for a Conversational Movie Recommendation System," in *Proceedings of the 7th International Conference on Human-Agent Interaction*, Kyoto Japan: ACM, Sep. 2019, pp. 135–143. doi: 10.1145/3349537.3351899.

[16] Y. Jin, L. Chen, W. Cai, and X. Zhao, "*CRS-Que*: A User-centric Evaluation Framework for Conversational Recommender Systems," *ACM Trans. Recomm. Syst.*, vol. 2, no. 1, pp. 1–34, Mar. 2024, doi: 10.1145/3631534.

[17] B. P. Knijnenburg, M. C. Willemsen, and A. Kobsa, "A pragmatic procedure to support the user-centric evaluation of recommender systems," in *Proceedings of the fifth ACM conference on Recommender systems*, Chicago Illinois USA: ACM, Oct. 2011, pp. 321–324. doi: 10.1145/2043932.2043993.

[18] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," *User Model. User-Adapt. Interact.*, vol. 22, no. 4–5, pp. 441–504, Oct. 2012, doi: 10.1007/s11257-011-9118-4.

[19] S. Fazeli, H. Drachsler, M. Bitter-Rijpkema, F. Brouns, W. van der Vegt, and P. B. Sloep, "User-centric evaluation of recommender systems in social learning platforms: accuracy is just the tip of the iceberg," *IEEE Trans. Learn. Technol.*, vol. 11, no. 3, pp. 294–306, 2017.

[20] L. W. Dietz, S. Myftija, and W. Wörndl, "Designing a Conversational Travel Recommender System Based on Data-Driven Destination Characterization.," in *RecTour@ RecSys*, 2019, pp. 17–21. Accessed: Dec. 02, 2024. [Online]. Available: http://www.ec.tuwien.ac.at/rectour2019/wp-content/uploads/2019/09/RecTour2019_Proceedings.pdf#page=24

[21] J. O. Álvarez Márquez and J. Ziegler, "Hootle+: A Group Recommender System Supporting Preference Negotiation," in *Collaboration and Technology*, vol. 9848, T. Yuizono, H. Ogata, U. Hoppe, and J. Vassileva, Eds., in Lecture Notes in Computer Science, vol. 9848. , Cham: Springer International Publishing, 2016, pp. 151–166. doi: 10.1007/978-3-319-44799-5_12.

[22] B. Loepp, T. Hussein, and J. Ziegler, "Choice-based preference elicitation for collaborative filtering recommender systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Toronto Ontario Canada: ACM, Apr. 2014, pp. 3085–3094. doi: 10.1145/2556288.2557069.

[23] E. C. Ling, I. Tussyadiah, A. Tuomi, J. Stienmetz, and A. Ioannou, "Factors influencing users' adoption and use of conversational agents: A systematic review," *Psychol. Mark.*, vol. 38, no. 7, pp. 1031–1051, Jul. 2021, doi: 10.1002/mar.21491.

[24] D. Gašević, V. Kovanović, and S. Joksimović, "Piecing the learning analytics puzzle: a consolidated model of a field of research and practice," *Learn. Res. Pract.*, vol. 3, no. 1, pp. 63–78, Jan. 2017, doi: 10.1080/23735082.2017.1286142.

[25] E. Fincham, D. Gašević, J. Jovanović, and A. Pardo, "From study tactics to learning strategies: An analytical method for extracting interpretable representations," *IEEE Trans. Learn. Technol.*, vol. 12, no. 1, pp. 59–72, 2018.

[26] M. Richardson and M. Healy, "Examining the ethical environment in higher education," *Br. Educ. Res. J.*, vol. 45, no. 6, pp. 1089–1104, Dec. 2019, doi: 10.1002/berj.3552.

[27] T. Cerratto Pargman and C. McGrath, “Mapping the ethics of learning analytics in higher education: A systematic literature review of empirical research,” *J. Learn. Anal.*, vol. 8, no. 2, pp. 123–139, 2021.
[28] S. Knight and S. B. Shum, “Theory and learning analytics,” *Handb. Learn. Anal.*, vol. 1, pp. 17–22, 2017.
[29] D. Gašević, S. Dawson, and G. Siemens, “Let’s not forget: Learning analytics are about learning,” *TechTrends*, vol. 59, no. 1, pp. 64–71, Jan. 2015, doi: 10.1007/s11528-014-0822-x.
[30] A. Woolfolk Hoy, H. A. Davis, and E. M. Anderman, “Theories of Learning and Teaching in *TIP*,” *Theory Pract.*, vol. 52, no. sup1, pp. 9–21, Oct. 2013, doi: 10.1080/00405841.2013.795437.
[31] L. Corsino and A. T. Fuller, “Educating for diversity, equity, and inclusion: A review of commonly used educational approaches,” *J. Clin. Transl. Sci.*, vol. 5, no. 1, p. e169, 2021.
[32] P. Jurado de los Santos, A.-J. Moreno-Guerrero, J.-A. Marín-Marín, and R. Soler Costa, “The Term Equity in Education: A Literature Review with Scientific Mapping in Web of Science,” *Int. J. Environ. Res. Public. Health*, vol. 17, no. 10, Art. no. 10, Jan. 2020, doi: 10.3390/ijerph17103526.
[33] M. Khalil, S. Slade, and P. Prinsloo, “Learning analytics in support of inclusiveness and disabled students: a systematic review,” *J. Comput. High. Educ.*, vol. 36, no. 1, pp. 202–219, Apr. 2024, doi: 10.1007/s12528-023-09363-4.
[34] C. Magno and J. Sembrano, “Integrating Learner Centeredness and Teacher Performance in a Framework,” *Int. J. Teach. Learn. High. Educ.*, vol. 21, no. 2, pp. 158–170, 2009.
[35] R. Collins, “Skills for the 21st Century: teaching higher-order thinking”.