

Explainable AI-Assisted Multi-Criteria Decision Support for Tutor Recommendation

Birat Aryal

Department of Software Engineering
Nepal College of Information Technology
Pokhara University
Kathmandu, Nepal
birat.221614@ncit.edu.np

Pranav Subedi

Department of Software Engineering
Nepal College of Information Technology
Pokhara University
Kathmandu, Nepal
pranav.221733@ncit.edu.np

Rasad Regmi

Department of Software Engineering
Nepal College of Information Technology
Pokhara University
Kathmandu, Nepal
rasad.221737@ncit.edu.np

Dipesh DC

Department of Software Engineering
Nepal College of Information Technology
Pokhara University
Kathmandu, Nepal
dipesh.221714@ncit.edu.np

Pranav Bhandari

Department of Software Engineering
Nepal College of Information Technology
Pokhara University
Kathmandu, Nepal
pranav.221732@ncit.edu.np

Abstract—Manual tutor assignment in online tutoring marketplaces becomes increasingly time-consuming, inconsistent, and challenging to scale as the number of available tutors grows. This paper proposes an Explainable AI-assisted Multi-Criteria Decision Support framework for tutor recommendation, developed and evaluated on the Hamro Tutor platform. The framework combines a deterministic Multi-Criteria Decision Making (MCDM) engine, which first applies grade-level, subject, and gender eligibility filtering and then ranks eligible candidate tutors using weighted scoring across location, availability, budget, qualification, and teaching experience, with a Large Language Model (LLM)-based explainability layer that generates human-readable justifications for each recommendation. Unlike fully autonomous recommendation systems, the proposed framework keeps the human administrator in the decision loop, using AI to support rather than replace human judgment. The system was evaluated on 10 real student tuition requests against a pool of 30 tutors, comparing manual and AI-assisted assignment workflows. Results show that AI-assisted selection reduced average decision time from 240.4 seconds to 22.6 seconds per request, a reduction of approximately 90.6%. Three platform administrators rated the suitability of AI-recommended tutors at an average of 4.57 out of 5, and rated the clarity of generated explanations at 4.77 out of 5. These findings suggest that explainable, human-in-the-loop AI decision support can meaningfully reduce administrative effort in tutor allocation while achieving positively perceived recommendation suitability and explanation quality.

Keywords—Explainable AI, Multi-Criteria Decision Making, Recommendation Systems, Human-in-the-Loop, Large Language Models, Tutor Matching, Decision Support Systems

I. INTRODUCTION

Marketplaces for online tutoring services have expanded greatly in the past few years, offering tutoring services for a variety of subjects to students at all levels. Websites such as Hamro Tutor use humans in the administration side of the process who connect incoming requests made by students to

appropriate tutors who are increasingly growing in numbers. At present, this connection process is fully manual whereby an administrator looks into the request made by the student, searches for the tutor database, screens candidates according to the requirements of the student and then makes the selection.

While this manual workflow is manageable when the number of registered tutors is small, it becomes increasingly time-consuming and difficult to scale as the tutor pool grows into the hundreds or thousands. Furthermore, because tutor selection depends heavily on individual administrator judgment, the process is susceptible to inconsistency and human bias, and offers no systematic mechanism for ensuring that recommendations are reproducible across administrators or over time.

Recommendation systems offer a natural solution to this scalability problem, and Multi-Criteria Decision Making (MCDM) techniques are particularly well suited to tutor-matching tasks, where a candidate's suitability depends on multiple, often competing, factors such as subject expertise, availability, budget, location, and experience. However, purely algorithmic recommendations present their own challenge: administrators are unlikely to trust or adopt a system whose decisions cannot be explained or justified in human-understandable terms, particularly in a context where the outcome affects a student's education.

This challenge motivates the integration of Explainable Artificial Intelligence (XAI) into the recommendation pipeline. Rather than presenting administrators with an opaque ranked list, the proposed system pairs a deterministic MCDM-based scoring engine with a Large Language Model (LLM)-based explanation layer that generates a concise, human-readable justification for each recommendation. Critically, the LLM is not involved in ranking or selecting tutors; its role is re-

stricted to explaining a decision already made by a transparent, reproducible scoring algorithm. This separation of concerns is intended to combine the reliability and reproducibility of a deterministic algorithm with the interpretability benefits of natural language explanation, while keeping the human administrator as the final decision-maker.

This paper presents the design and evaluation of this Explainable AI-assisted Multi-Criteria Decision Support framework, applied to real tutor allocation data from the Hamro Tutor platform. Specifically, this work investigates the following research questions:

- **RQ1:** Can an MCDM-based recommendation engine reduce tutor selection time compared to manual assignment?
- **RQ2:** Do administrators perceive AI-generated explanations as clear and useful?
- **RQ3:** Do administrators perceive AI-recommended tutors as well-aligned with sound tutor-matching judgment for a given student's requirements?

The main contributions of this research are summarized as follows:

- A tutoring recommendation system based on human judgment, that uses a rational process for decision making as well as explanations generated by a large language model.
- A transparent weight assignment mechanism, that is employed after level-based, subject-based, and gender-based filtering as well as taking into consideration factors such as location, time, budget, qualifications, and experience of candidates.
- A real world study of the effectiveness of AI tutors in selecting suitable tutors for the students through data analysis from an online platform.

The rest of this paper is structured as follows. Section II is devoted to a review of previous studies in explainable recommendation systems, tutor and teacher recommendation, multi-criteria decision-making, and LLM-based recommendation explanation. In Section III, the framework and system architecture are presented. The methodological approach used in the experiments is described in Section IV. Section V provides experimental results. Discussion and implication of our study are provided in Section VI, and Section VII and Section VIII conclude this paper.

II. RELATED WORK

A. Explainable Recommendation and Trust Calibration

Naiseh *et al.* [1] explore explainability in Human-AI collaborative decision-making to better calibrate trust. Their think-aloud studies and co-design sessions with medical practitioners reveal behavior such as skipping and misapplying explanations and discuss design considerations including engagement, attention guidance, friction, and support for learning. A main conclusion is that explanations should be tailored to interaction and decision context rather than being presented as standalone. However, their work focuses on medical decision support and

does not consider deterministic Multi-Criteria Decision Making (MCDM), tutor recommendation, or LLM-based natural language explanations.

B. Explanation Evaluation, Faithfulness, and Scrutability

Chen *et al.* [2] presented a survey on explainable recommendation evaluation, which summarized over 100 related works and categorized the evaluated approaches from four perspectives: effectiveness, transparency, persuasiveness, and scrutability. They also grouped the evaluation methods into case studies, quantitative evaluation, crowdsourcing, and online experiments and discussed their pros and cons. In particular, their survey showed that evaluation of explanation effectiveness could involve whether explanations enable users to make decisions faster or more accurately, while scrutability focused on the explanation-result correspondence. This survey offers a methodological basis for assessing the quality of explanations according to their decision-support objective.

Xu *et al.* [6] introduced a framework for explainable recommendations that takes into account the aspect of explanation faithfulness and system scrutability in a general reusable framework. The framework introduced by Xu *et al.* [6] uses two separate models for recommendation and explanation using knowledge distillation style training to learn decision rules from the recommendation model. They experimentally demonstrated better faithfulness in their proposed framework compared to other methods along with providing support for scrutinizing explanations. Their work points out that the evaluation of post-hoc explanations must consider faithfulness to the recommendation system itself.

C. Tutor and Teacher Recommendation

Ach'on *et al.* [3] developed 'SOS TUTORÍA UC', a tutor recommendation system that aims at pairing students with appropriate peer tutors based on academic aptitude, personality attributes, and geographic proximity. This system uses the Big Five personality dimensions and the WeNet platform for recommending tutors based on users' preference of similarity, dissimilarity, or indifference of the users' personality type. A pilot test conducted on 43 participants showed positive results in general, though there were some recommendations that did not entirely fulfill the user's criteria and need for improvement. But the system neither specifies an MCDM-based model nor provides any LLM-generated natural language explanation.

A course-teacher recommendation system using a combination of teaching style and TPACK (Technological Pedagogical Content Knowledge), along with the difficulty of the course and teaching quality, was designed by Yao *et al.* [7]. The model uses an advanced weighted bipartite graph, Personal-Rank, and Slope One algorithm to handle sparsity and cold start issues prior to creating TOP-N teacher recommendations. Experiments conducted on three datasets of different university courses found the method superior in terms of Precision, Recall, and RMSE scores when compared with various other collaborative filtering models and graph-based systems. Nevertheless, the method is specific for universities' course-teacher

allocation and ignores practical constraints such as budget, location, availability, and qualifications of the tutor in the marketplace.

D. Multi-Criteria Decision Making

Multi-Criteria Decision Making (MCDM) methods have been widely applied in recommendation problems where decisions depend on multiple conflicting criteria. Techniques such as Simple Additive Weighting (SAW) and TOPSIS have been used to rank alternatives based on user-defined preferences, with SAW in particular offering a transparent, computationally lightweight aggregation approach well suited to systems that require interpretable, auditable scoring. Compared with black-box machine learning approaches, MCDM provides transparent decision rules and allows stakeholders to understand how individual criteria influence final rankings. Kaliszewski and Podkopaev [4] examined Simple Additive Weighting (SAW) as an explanatory framework for Multiple Criteria Decision Analysis methods, showing how weighted linear functions over normalized criteria can be used to interpret and, when feasible, reproduce MCDA rankings. Their analysis demonstrates that feasible sets of criterion weights can preserve a given ranking and can be used to examine the influence and stability of criteria weights. However, existing MCDM-based recommendation approaches generally lack natural-language explanation capabilities, limiting their usability in domains requiring high user trust, such as education.

E. LLM-Based Recommendation Explanations and Research Gap

Ma *et al.* [5] proposed XRec, a model-agnostic framework that integrates graph-based collaborative filtering with Large Language Models (LLMs) to generate textual explanations for recommendation outputs. XRec uses LightGCN to capture collaborative user-item relationships and a lightweight adapter to align these representations with an LLM for explanation generation. Experiments on Amazon Books, Yelp, and Google Reviews showed that XRec achieved strong performance against several baseline explanation-generation methods across semantic explainability and stability metrics. The study demonstrates the potential of LLMs for generating personalized recommendation explanations, but its collaborative-filtering setting differs from the explicitly weighted Multi-Criteria Decision Making (MCDM) and human-administered tutor allocation considered in this work.

Qin *et al.* [8] investigated the credibility of explanations in LLM-based conversational recommender systems, noting that highly persuasive language models can generate explanations containing misleading or factually inconsistent information. They proposed PC-CRS, which combines credibility-aware persuasive strategies with iterative self-reflective refinement to identify and remove unsupported information from generated explanations. Experiments on the ReDial and OpenDialKG benchmarks showed improvements in explanation credibility and persuasiveness over several baseline conversational recommender systems. However, the study focuses on conversational

recommendation and does not address explicitly weighted Multi-Criteria Decision Making (MCDM) for tutor ranking or administrator-assisted professional tutor allocation.

Collectively, this body of work establishes that explanation quality must be evaluated through human-centered criteria such as trust, transparency, and usefulness [1], [2], that explanation faithfulness and credibility cannot be assumed and require explicit evaluation [6], [8], and that AI-assisted tutor and teacher matching can incorporate multiple structured attributes while preserving human decision authority [3], [7]. However, our review identified limited prior work combining an explicit, transparently weighted MCDM ranking algorithm with LLM-generated natural language explanations, nor do these studies address administrator-assisted professional tutor allocation in educational marketplaces, where constraints such as budget, location, qualifications, and availability must be jointly balanced. Motivated by these gaps, the proposed research develops an Explainable AI-Assisted Multi-Criteria Tutor Recommendation System that integrates a deterministic MCDM-based weighted scoring engine with a Gemini-generated explanation layer, evaluated using both objective performance metrics and human-centered assessment criteria, within a human-in-the-loop decision support paradigm for professional tutor allocation.

III. PROPOSED FRAMEWORK

A. System Architecture

The proposed framework is structured as a human-in-the-loop, AI-assisted decision support system rather than a fully autonomous recommendation engine. As illustrated in Fig. 1, the system processes an incoming student request through a deterministic Multi-Criteria Decision Making (MCDM) engine, which evaluates all candidate tutors and produces a ranked recommendation. This ranked list, together with the tutor profiles and the original student request, is then passed to a Large Language Model (LLM)-based explainability layer, which generates a concise natural-language justification. The administrator reviews both the recommendation and its explanation before acting on it, ensuring that the final tutor-selection decision remains with the administrator.

The framework operates in four stages. First, a student submits a tuition request specifying subject, grade level, budget, location, and preferred schedule. Second, the MCDM engine evaluates every tutor in the platform's database against this request, first applying eligibility filtering and then computing a weighted suitability score for each eligible candidate. Third, the highest-ranked tutors (top five) are passed to the Gemini API, which generates a structured explanation describing why the highest-ranked tutors are suitable, grounded in the specific criteria that contributed to the match. Fourth, the administrator reviews the recommended tutor and accompanying explanation through the administrative dashboard and then decides whether to act on the recommendation or select an alternative tutor, after which the standard contact and demo-class workflow continues unchanged.

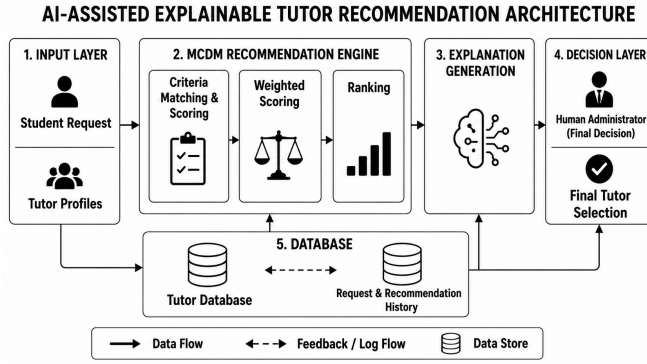


Fig. 1. Architecture of the proposed Explainable AI-assisted tutor recommendation framework.

A central design principle of the proposed framework is the strict separation between the ranking function and the explanation function. Tutor ranking is performed exclusively by the deterministic MCDM scoring algorithm, ensuring that recommendations are reproducible, auditable, and free from the hallucination risk associated with LLM-based decision-making. The LLM is deliberately restricted to a non-decisional role: generating a human-readable explanation for a ranking that has already been determined. This design choice reflects the recommendation, drawn from the explainable decision-support literature discussed in Section II, that transparency and reliability are essential for user trust in AI-assisted decision contexts, particularly where outcomes affect real students' educational experiences.

The architecture separates the recommendation and explanation processes to ensure that the LLM does not influence the ranking decision. This separation improves reproducibility, reduces the potential for hallucination to affect ranking decisions, and allows administrators to verify recommendations before final assignment.

B. MCDM Recommendation Engine

Before scoring, each candidate tutor is evaluated against three hard eligibility filters: grade-level compatibility (the tutor must be qualified to teach at or above the student's requested grade level), subject match (the tutor must teach at least one of the requested subjects), and gender preference (applied only when the student specifies one). Tutors that fail any of these filters are excluded from ranking; the remaining eligible tutors proceed to the weighted-scoring stage.

Tutor suitability among eligible candidates is computed as a weighted linear aggregation across five criteria: location match, availability match, budget match, qualification level, and teaching experience. Formally, the suitability score of each eligible candidate tutor is given by

$$Score(T_i) = \sum_{j=1}^n w_j C_{ij} \quad (1)$$

where T_i represents a candidate tutor, $C_{ij} \in [0, 1]$ represents the normalized performance score of tutor i on criterion j ,

TABLE I
WEIGHTED CRITERIA AND ASSIGNED WEIGHTS

Criterion	Weight (points)
Location compatibility	25
Availability matching	25
Budget compatibility	15
Qualification level	15
Teaching experience	10
Total	90

and w_j represents the maximum achievable points (weight) for criterion j , with $n = 5$ weighted criteria. The maximum possible total score is therefore $\sum_{j=1}^5 w_j = 90$. Table I lists the five criteria and their assigned weights.

The weights were assigned based on the research team's direct operational experience administering tutor-student matching on the Hamro Tutor platform: location and availability were identified as the two most frequent practical blockers to a successful match and were therefore weighted most heavily (25 points each), followed by budget and qualification (15 points each) and teaching experience (10 points). These weights reflect domain-informed operational priorities and were not derived through an automated optimization procedure; they were also not subjected to sensitivity analysis, and can be revised as additional assignment data accumulates.

Each criterion is scored according to a defined rule. For example, the location criterion awards 25 points for an exact locality match, 22 for a partial textual match, 18 when tutor and student share the same city prefix, and 4 when localities differ. Budget compatibility awards 15 points when the tutor's fee is equal to the student's budget, 12 when the relative difference does not exceed 10%, 8 when it does not exceed 25%, and 0 otherwise; a negotiable fee is treated as fully compatible. Schedule availability combines a day-level Jaccard similarity (capped at 15 points) with a time-overlap component (up to 10 points). Qualification is scored by academic rank (doctorate or professional teaching license rated highest), and teaching experience is scored in increasing brackets up to 10 points.

Each eligible tutor is scored across the five weighted criteria and combined into a final score using Equation 1. Tutors are ranked in descending order of total score, and the highest-ranked tutors are passed to the explainability module for justification generation.

C. Explainability Module

The highest-ranked tutors (top five), together with their per-criterion score breakdowns and the corresponding student request, are passed to an LLM-based explainability module built on the Google Gemini API, with `gemini-3.1-flash-lite` as the default model and a fallback to a more recent Flash-Lite variant when required. The prompt explicitly instructs the model that it must not rank or re-order tutors — the ranking has already been computed by the deterministic scoring algorithm — and constrains the output to a concise structured explanation of fewer than 350

words, organized under four required headings: Recommended Tutor, Alternative Options, Things to Consider, and Overall Recommendation. The model is instructed to ground its justification in the criteria and values actually provided and to avoid inventing facts not present in the input. The LLM plays no role in ranking or selection; its function is limited to explaining a decision already determined by the scoring algorithm.

IV. EXPERIMENTAL METHODOLOGY

A. Dataset

Hamro Tutor (hamrotutor.tech) is an online tutoring marketplace platform where students submit tuition requirements and administrators manually assign suitable tutors based on multiple constraints. Evaluation used a pool of 30 tutors from the Hamro Tutor database and 10 real student tuition requests (S1–S10), each specifying subject, grade, budget, location, and preferred schedule. The same 10 requests were used across both manual and AI-assisted conditions to enable a direct comparison of the two workflows.

B. Evaluation Setup

Two evaluations were conducted. First, a decision time comparison: for each request, time-to-selection was recorded under manual selection (administrator searching the full 30-tutor pool without assistance) and AI-assisted selection (administrator reviewing the system's top recommendation and explanation before acting). Second, a quality evaluation: three platform administrators independently rated each of the 10 recommendations on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) across two dimensions suitability of the recommended tutor for the student's requirements, and explainability, i.e., whether the generated explanation clearly justified the recommendation.

Both manual and AI-assisted decision times were recorded directly by the research team while operating the platform, using the same ten requests in the same order under each condition; no independent randomization or counterbalancing of request order was applied. The three evaluating administrators are members of the research team and co-administrators of the Hamro Tutor platform, responsible for its day-to-day tutor-matching operations. This provided evaluators with direct domain expertise in tutor-student matching, but also means they were not independent of the system under evaluation and were aware of which condition (manual or AI-assisted) each recommendation came from.

C. Evaluation Metrics

The proposed framework is evaluated using three metrics.

Decision Time Reduction: measures the reduction in decision time achieved by comparing manual and AI-assisted tutor selection.

Recommendation Quality: measures whether administrators consider recommended tutors suitable for student requirements using a five-point Likert scale.

Explanation Quality: measures the clarity and usefulness of generated explanations based on administrator evaluation.

TABLE II
DECISION TIME COMPARISON BETWEEN MANUAL AND AI-ASSISTED TUTOR SELECTION

Request	Manual (s)	AI-Assisted (s)	Time Saved (s)
S1	248	24	224
S2	231	22	209
S3	265	26	239
S4	219	20	199
S5	242	23	219
S6	236	21	215
S7	254	25	229
S8	227	20	207
S9	244	23	221
S10	238	22	216
Average	240.4	22.6	217.8

V. RESULTS

This section presents the evaluation results of the proposed framework in terms of decision-making efficiency, recommendation quality, and explanation clarity.

A. Decision Time Comparison

In Table II, the time needed to perform the tutor selection in both manual and AI-based modes is contrasted. The average time of tutor selection through the manual method amounted to 240.4 seconds for all ten students, and the proposed approach made the average decision time equal to 22.6 seconds. Thus, there was a decrease of 217.8 seconds on average, i.e., nearly 90.6% of the decision time was decreased.

As one can see, the decrease in the time of processing was consistently observed for all assessed requests with time savings of 199 to 239 seconds each. These results indicate that the proposed framework substantially decreases the administrative effort required for tutor selection while maintaining a consistent recommendation process.

B. Recommendation Quality

The appropriateness of the recommended tutors was independently assessed by three senior administrators of the platform using a five-point Likert scale, where 1 represents Strongly Disagree and 5 represents Strongly Agree. An assessment was made by each administrator as to whether the tutor recommended by the proposed approach is an appropriate match for the student request under consideration.

According to Table III, the proposed framework managed to achieve an overall mean recommendation quality of 4.57 out of 5. All assessed requests got an average score of at least 4.00, which demonstrates agreement among the administrators that the recommended tutors were appropriate. Some of the requests had full agreement among the assessors, which suggests high agreement regarding the appropriateness of the recommendations generated by the MCDM-based ranking approach.

C. Explainability Evaluation

The effectiveness of the explanation component was determined by the same group of administrators who assessed how clearly the LLM-generated explanations were able to

TABLE III
ADMINISTRATOR EVALUATION OF RECOMMENDATION QUALITY

Request	Admin 1	Admin 2	Admin 3	Mean
S1	5	5	4	4.67
S2	4	5	4	4.33
S3	5	5	5	5.00
S4	4	4	5	4.33
S5	5	4	5	4.67
S6	4	5	4	4.33
S7	5	5	5	5.00
S8	4	4	4	4.00
S9	5	5	4	4.67
S10	5	4	5	4.67
Overall Mean				4.57

TABLE IV
ADMINISTRATOR EVALUATION OF EXPLAINABILITY

Request	Admin 1	Admin 2	Admin 3	Mean
S1	5	5	5	5.00
S2	5	4	5	4.67
S3	5	5	5	5.00
S4	4	5	5	4.67
S5	5	5	4	4.67
S6	5	4	5	4.67
S7	5	5	5	5.00
S8	4	4	5	4.33
S9	5	5	5	5.00
S10	5	5	4	4.67
Overall Mean				4.77

explain the reason for the recommendation of the tutors. The explanations were assessed through the use of the five-point Likert scale in terms of clarity, relevance, and usefulness.

As shown in Table IV, the explainability component managed to obtain a mean score of 4.77 out of 5, meaning that the administrators were able to understand the generated explanations clearly. Most of the explanations scored either 4 or 5, indicating the success of the natural language justifications.

D. Summary of Results

In conclusion, the overall assessment proves that the framework significantly cuts down tutor selection time and generates appropriate, clear, and usable recommendations and explanations to meet its desired purpose of tutor recommendation efficiently, clearly, and human-centrally.

VI. DISCUSSION

The findings support RQ1: the proposed MCDM-based recommendation engine cut down tutor selection time by around 90.6%. For RQ3, the high quality of the recommendations ratings (4.57 out of 5) suggests that the administrators found the recommendations of AI to be in agreement with the good judgment on tutor-matching. Finally, for RQ2, the administrators rated the generated explanations as clear and valuable (mean 4.77 out of 5). However, since no no-explanation condition was tested in this experiment, it is impossible to conclude that the explanation generation actually led to increased trust. Still, the fact that the generated explanations were rated highly means that the condition for increased trust was satisfied, albeit not sufficient by itself.

Such a dramatic improvement in decision time suggests that the framework can significantly streamline tutor assignment by performing the initial evaluation of tutor profiles, thereby reducing the time required for tutor selection. The administrator no longer has to review multiple tutor profiles for each new student request. Instead, the administrator can review the recommendation generated by the AI-assisted system together with its accompanying explanation.

The quality of recommendations also shows that the selected decision criteria are a good representation of the practical factors that are considered during tutor assignment. Applying filters by grade level, subject, and gender, followed by evaluating the location, availability, budget compatibility, teaching experience, and educational qualifications through a weighted scoring system, the deterministic MCDM engine generates recommendations that are always perceived as reasonable by the evaluators. The good agreement between evaluators suggests that the proposed scoring method works well with the existing tutor selection practices.

The explainability component was positively perceived as part of the overall decision support process. In contrast to using the LLM for generating recommendations, the proposed approach limits the role of the LLM to providing an explanation of decisions that were made by the deterministic ranking engine. Such design increases explainability without compromising the reproducibility of the recommendation process. The administrators can see the reasons for recommending the tutor without knowing the details about the scoring system, which could be helpful in increasing administrators' trust to the system, even though such an aspect was not examined in this study.

In comparison to previously used tutor recommendation frameworks, this approach has two more benefits. The first is that the MCDM module makes ranking decisions openly, unlike the other methods using opaque prediction models. The second benefit is that the LLM-powered interpretation layer helps make numerical ranking values interpretable to the administrator.

VII. LIMITATIONS AND FUTURE WORK

Despite these encouraging results, several limitations should be acknowledged. The evaluation was conducted using a limited number of student requests (10) and administrator evaluators (3) within a single educational marketplace. Consequently, the findings may not fully generalize to larger datasets, different geographical regions, or other tutoring platforms. In addition, recommendation quality was assessed through administrator perception rather than long-term educational outcomes such as student satisfaction, learning performance, or tutor retention.

The evaluation itself was conducted externally by the research team while operating the platform, rather than through automated instrumentation embedded in the system; timing measurements and ratings were recorded manually, and the evaluating administrators were members of the research team responsible for operating the platform. Their assessments may

therefore be subject to evaluator bias, and future work should incorporate independent or blinded evaluators to validate these findings. Similarly, because decision times were recorded using a fixed request order without randomization or counterbalancing, order effects (e.g., administrator familiarity with recurring requests) cannot be ruled out as a contributing factor to the observed time reduction.

The study did not include a no-explanation control condition, so a causal link between explanation quality and administrator trust cannot be established; the results support the more limited conclusion that explanations were perceived as clear and useful. Furthermore, the MCDM weights were assigned based on operational experience and were not empirically optimized or subjected to sensitivity analysis; investigating the stability of rankings under alternative weight configurations is an important direction for future work. The LLM-based explanation layer also has limitations: explanations are generated with a non-deterministic model (with fallback between Flash-Lite variants), outputs are not cached across sessions, and no automatic runtime validation currently verifies that every statement in an explanation matches the computed scores; explicit faithfulness evaluation of the explanation layer remains open work.

Future work will focus on evaluating the framework using larger and more diverse datasets, incorporating additional recommendation criteria where appropriate, performing sensitivity analysis on the scoring weights, and developing an automated faithfulness-validation layer for generated explanations. Furthermore, future studies may compare the proposed deterministic MCDM approach with machine learning-based recommendation models while maintaining explainability and human oversight as core design principles.

VIII. CONCLUSION

In conclusion, this paper presented an explainable AI-assisted tutor recommendation framework that aims at supporting administrative decision-making in educational marketplaces. In particular, the framework proposes the use of a deterministic MCDM engine for ranking tutors and an LLM-based explainability module for generating justifications for the given recommendations. In contrast to existing architectures, the proposed one separates the ranking step from explanation generation to ensure transparency and auditability of the recommendations as well as interpretable outputs for administrators.

The performance of the proposed framework was evaluated based on typical tutor assignment requests to examine the improvements in decision-making efficiency, quality of recommendations, and clarity of explanations. The obtained experimental results show that there was a considerable decrease in tutor selection time while maintaining high administrator ratings for recommendation suitability and explanation quality. These results suggest that the proposed solution can improve the process of assigning tutors without removing human decisions and supporting the human-in-the-loop decision making

paradigm where the final tutor-selection decision is still with the administrator.

Therefore, the main contribution of this research is the development of an explainable decision support framework combining deterministic multi-criteria evaluation and LLM-based explanations for tutor recommendation. Different from other architectures, the suggested one pays.

REFERENCES

- [1] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "Explainable Recommendation: When Design Meets Trust Calibration," *World Wide Web*, vol. 24, pp. 1857–1884, 2021, doi: 10.1007/s11280-021-00916-0.
- [2] X. Chen, Y. Zhang, and J.-R. Wen, "Measuring 'Why' in Recommender Systems: A Comprehensive Survey on the Evaluation of Explainable Recommendation," arXiv preprint arXiv:2202.06466, 2022.
- [3] L. Achón, A. de Souza, A. Hume, R. Chenu-Abente, A. de Götzen, and L. Cernuzzi, "SOS TUTORÍA UC: A Diversity-Aware Application for Tutor Recommendation Based on Competence and Personality" (Short Paper), in *Proc. Int. Conf. on Hybrid Human-Artificial Intelligence (HHAI 2023)*, 2023, pp. 243–247.
- [4] I. Kaliszewski and D. Podkopaev, "Simple additive weighting—A metamodel for multiple criteria decision analysis methods," *Expert Systems with Applications*, vol. 54, pp. 155–161, 2016, doi: 10.1016/j.eswa.2016.01.042.
- [5] Q. Ma, X. Ren, and C. Huang, "XRec: Large Language Models for Explainable Recommendation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 391–402, doi: 10.18653/v1/2024.findings-emnlp.22.
- [6] Z. Xu, H. Zeng, J. Tan, Z. Fu, Y. Zhang, and Q. Ai, "A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability," *ACM Transactions on Information Systems*, vol. 42, no. 1, Art. no. 29, pp. 1–29, 2023, doi: 10.1145/3605357.
- [7] D. Yao, X. Deng, and X. Qing, "A Course Teacher Recommendation Method Based on an Improved Weighted Bipartite Graph and Slope One," *IEEE Access*, vol. 10, pp. 129763–129780, 2022, doi: 10.1109/ACCESS.2022.3228957.
- [8] P. Qin, C. Huang, Y. Deng, W. Lei, and T.-S. Chua, "Beyond Persuasion: Towards Conversational Recommender System with Credible Explanations," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 4264–4282, doi: 10.18653/v1/2024.findings-emnlp.247.