Improving Course Recommendation Systems with Explainable AI: LLM-Based Frameworks and Evaluations

Jiawei Li
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
jiawei009@e.ntu.edu.sg

Qianru Lyu
National Institute of Education
Nanyang Technological University, Singapore
nie20.lq@e.ntu.edu.sg

Wei Qiu
Centre for the Applications of Teaching and Learning Analytics for Students (ATLAS)
Nanyang Technological University, Singapore
qiuwei@ntu.edu.sg

Andy W. H. Khong
School of Electrical and Electronic Engineering
Lee Kong Chian School of Medicine
Nanyang Technological University, Singapore
andykhong@ntu.edu.sg

ABSTRACT

Deep learning-based course recommendation systems often suffer from a lack of interpretability, limiting their practical utility for students and academic advisors. To address this challenge, we propose a modular, post-hoc explanation framework leveraging Large Language Models (LLMs) to enhance the transparency of deep learning-driven recommenders. Our approach utilizes course descriptions, social science theories, and structured explanation formats to generate human-readable justifications, improving the interpretability and trustworthiness of recommendations. This study aims to enhance the AI-generated course recommendations by empirically evaluating the different LLM-based explanations for course recommendations. With the proposed explanation generation pipeline, four LLM-based explanations were generated and surveys were collected from course instructors to understand the efficiency of each prompt design. Evaluation with three instructors indicates that prompts integrating course context and the theory of relevance significantly enhance explanation quality and user satisfaction. Our findings highlight the importance of content-specific elements in interpretable AI-driven educational tools, with implications for enhancing explainability in learning analytics. This study provides insights for future fine-tuning of course recommendation systems supported by explainable artificial intelligence (XAI).

Keywords

Explainable artificial intelligence (XAI), course recommendation system, large language models (LLM), explanation generation pipeline, explanation evaluation

Jiawei Li, Qianru Lyu, Wei Qiu, and Andy W. H. Khong. Improving Course Recommendation Systems with Explainable AI: LLM-Based Frameworks and Evaluations. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 205–214. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870185

1. INTRODUCTION

Artificial intelligence (AI) systems have been widely applied in e-learning systems to provide personalized learning experiences for enhanced engagement and achievement of academic learning outcomes [20, 19]. These systems have the ability to predict academic performance [13, 22] and recommend suitable courses according to a set of criteria [15, 11]. However, stakeholders (students, educators, etc.) often lack domain-specific experience and face challenges in interpreting and optimizing these AI systems. To this end, explainability plays a critical role in AI in education—knowledge of how a learner achieves their learning outcome provides context and guides the decision-making process [12].

The importance of explainability in educational AI systems can be described from several perspectives. First, in addition to increasing the adoption rate, explanations can result in actionable interventions for students, leading to better learning experiences and achievements of learning outcomes. For example, the performance of deep learning models can be enhanced by providing the rationale behind the recommendations [23]. Furthermore, explainable AI (XAI) tools facilitate insights leading to better educational practices that improve student support to achieve learning goals [2].

XAI tools embedded within course recommendation systems can assist learners navigate their academic pathways and provide personalized course suggestions according to their learning objectives [18, 15]. These systems are designed to guide students through the complex course catalog (which is often scaffolded) and help them make informed decisions about which courses to read [16]. Previous studies have proposed rule-based learning in course recommendation systems by providing a chain of steps through well-organised rules, while simultaneously avoiding overly complex rules that can lower the quality of recommendations [7]. These rule-based systems, however, cannot adapt to changes within the curriculum leading to erroneous recommendations.

The extraction of keywords from course descriptions has also

been exploited for course recommendation[28]. According to Hilton's theory of relevance [9], explanations are more effective when causal relationships are explored within a given context. However, such keyword-based method does not provide conversational explanations nor do they fully consider a student's academic profile or the deeper contextual relevance of course content. Despite growing efforts to provide effective course recommendations, it is still unclear which explanation design can better support university course recommendation systems. The challenge of generating effective, interpretable explanations that genuinely assist students in the course selection process remains unresolved.

In the recent decade, Large Language Models (LLMs) have demonstrated considerable potential in the domain of generating explanations for AI systems. LLMs possess the ability to process and generate natural language text, making them suitable for providing comprehensible and context-specific explanations [27]. Attempts have been made to develop analogous educational applications that leverage LLM to provide explanations to users. For example, Swamy et al. have proposed the LLM-based iLLuMinaTE pipeline, which aims to provide explanations to student performance feedback and enhance its usability and actionability for students [24]. Similarly, an LLM-based chatbot has been employed for learning path recommendation, yielding satisfactory user ratings for the generated responses. These efforts demonstrate the potential of LLMs in addressing the interpretability challenges currently faced by course recommendation systems [1].

With growing evidence that LLMs can improve the explainability of AI-based course recommendations, the evaluation and optimization of LLM prompt designs pose a critical challenge. Although automatic metrics such as BLEU [17] and ROUGE [14] allow rapid evaluation of the design of LLM prompts by evaluating the generated text, their reliance on ground-truth references makes them less suitable for certain educational applications, including explanatory course recommendation systems. Consequently, human evaluation remains indispensable as it provides valuable insights into the alignment of generated text with human expectations and its practical applicability [25]. Surveys evaluating user experience are a crucial component in assessing the effectiveness of XAI in educational settings [4]. To fine-tune AI-generated course recommendations, an in-depth evaluation of different prompt designs from stakeholders' perspectives is necessary.

In educational applications, diverse cognitive theories have been integrated into prompts to enhance the utility and actionability of LLM-generated explanations for students [24]. Similarly, different LLM models have also been compared according to their ability to assist in the instruction of an undergraduate statistics course [21]. Given that prompts can significantly influence the generation process, these studies underscore the need to examine the factors that influence the generation of explanations to optimize educational outcomes in course recommendation systems. In particular, actors such as patterns that translate different knowledge into software and structuring prompts profoundly affect the quality of the generated text. The manner in which course descriptions are incorporated into prompts may also impact the effectiveness of explanations, necessitating further exploration. It is useful to note that outcome-based descriptions [8] are often based on a set of predefined learning outcomes, whereas the content-based descriptions [6] involve the description of topics and subject areas, which, to a large extent, are related across different courses. As discussed in [24], incorporating theories from the social sciences can also substantially improve the quality of AI-generated explanations in educational settings.

To address the gaps above, this work improves AI-generated course recommendations by empirically evaluating the different LLM-based explanations for course recommendations. We propose a two-stage, post-hoc explainable course recommendation system framework that leverages course descriptions. In the first stage, a machine learning-based recommendation model generates a list of suggested courses for students. The second stage provides interpretability by generating explanations that assess the suitability of the recommended courses with rationale being guided by the proposed explanation generation pipeline. Using a structured prompt design approach, we evaluated four distinct prompts, each incorporating unique prompt patterns, course description focuses, and applications of social science theories. We assess the effectiveness of the proposed approach by conducting a study in which courses within a university were recommended based on students' prior academic performance and, thereafter, with corresponding explanations generated. The evaluation was performed by subject-domain experts and learning science researchers, who evaluated the quality of the generated explanations by comparing them against a set of predefined criteria. Likert scales were used to quantify their evaluations along with qualitative feedback to highlight the strengths and weaknesses of the explanations. Three instructors with different levels of experience (in terms of years and courses taught) evaluated the explanations associated with six students with diverse academic backgrounds. Ratings were based on multiple scoring criteria across 120 explanations. Analysis of these evaluations offers valuable insight into the overall effectiveness and interpretability of the explanations generated by different prompts.

In this study, two research questions will be answered:

1. What is the efficiency of the proposed XAI course recommendation system framework in the university context?
2. What are the key factors that contribute to effective explanations in educational contexts?

2. METHODOLOGY

The overall workflow of the proposed framework is illustrated in Figure 1, which consists of three main components: the course recommendation module, the explain module, and expert evaluation.

2.1 Participants and learning context

This study was conducted at an engineering school in a university. In the explanation generation phase, the academic profile of six undergraduate engineering students were randomly selected each category—top, average, and struggling performers. In the explanation evaluation phase, three course instructors from the same university's engineering

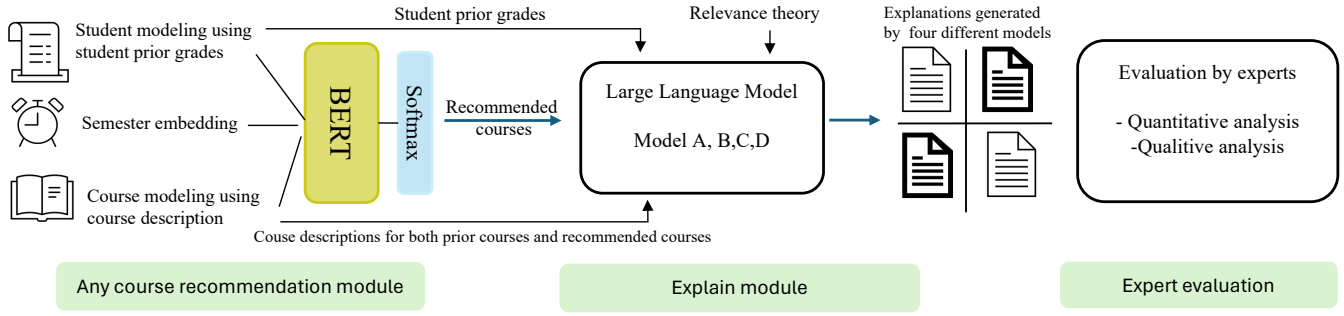


Figure 1: Overall workflow of the explainable course recommendation framework.

school participated. These instructors consisted of experienced course coordinators, a senior lecturer, and a junior lecturer, all of whom are responsible for teaching engineering courses. The instructors had varying levels of teaching experience and curriculum design.

2.2 Explanation module development

The input to the explanation module is a list of recommended courses generated by the course recommendation module. The explanation module of the proposed framework generates explanations detailing why these recommendations were generated using various strategies. These strategies adopt common prompt patterns and integrate social science and learning theories for enhancing their effectiveness by generating explanations that are informative and pedagogically effective.

Generation of recommendation list. The objective of the recommendation module is to generate a ranked list of courses for students, a task that can be accomplished using any course recommendation model. In this work, the recommended courses are generated using a Bidirectional Encoder Representations from Transformers (BERT)-based course recommendation model [5]. The recommendation module uses both historical academic achievements and course descriptions. These course descriptions include the specific topics covered in chronological order and are provided by the course coordinators for the university's approval. Hence, student performance embeddings are defined based on prior grades achieved over the past semesters, while course embeddings are generated using an LLM-based embedding function derived from course descriptions. The input representation of BERT consists of a sequence of tokens processed into embeddings that include: token embeddings, segment embeddings, and position embeddings. Similarly, student embeddings, course embeddings, and a one-hot semester embedding that indicates the semester in which a course is taken are generated. The BERT model's output is processed by a softmax function, which generates a ranked list of recommended courses, with higher-ranking courses having a greater likelihood of being recommended.

Explanation Generation Pipeline. The explanation generation pipeline in our framework consists of four different parts shown in Figure 2:

- General Introduction: presents an overview of the recommended course and its relevance to the student.

General Introduction: *You are an **advisor for university students in terms of their course selection**. Based on the courses the student taken before and the courses we recommend, please evaluate ...*

Student Prior Performance:
e.g. Course ID: Effective communication, Grade: 70.0 ...

Course description:
***Outcome-based:** The aim of this course is hence to take students with no prior experience of thinking...*
***Content-based**: Unit 1: Introduction to communication*
Unit 2: Writing short proposals...

Detailed requirement:
e.g. The explanation should:
** **Identify relevant skills or knowledge** acquired in the student's previous courses that are **directly applicable** to the recommended course...*

Figure 2: Explanation generation pipeline.

- Student Prior Academic Performance: highlights the student's academic performance across past semesters and how it relates to the recommended course.
- Course Descriptions: detailed descriptions of the recommended course, including learning outcomes and content.
- Detailed Requirements and Constraints: specific requirements and constraints that the student should consider when enrolling in the course.

Model Variations. To generate relevant explanations that support decision making, there exist several factors affecting the quality of generation:

1. **Prompting pattern:** A prompt is a set of instructions that guides an LLM to generate responses for a specific task. In our context of explanation generation for effective course recommendation, we consider two widely accepted prompt patterns: *persona pattern* and *cognitive verifier pattern*. Persona pattern directs the

Table 1: Comparison of Explanation Models

Model	Prompting Pattern	Course Description	Relevance Theory
A	Persona	Outcome-Based	No
B	Cognitive Verifier	Outcome-Based	No
C	Cognitive Verifier	Outcome-Based	Yes
D	Cognitive Verifier	Content-Based	Yes

Table 2: Sample explanations generated by different model

Model	Sample
A	The student has completed the foundation course CM001 in communication. CM002 is an advanced course building upon those skills. Given the student's prior successful completion of the foundational course, this advanced course is suitable. The student's high grade in CM001 also indicates a good foundation and aptitude for this subject.
B	The student has a strong background in communication skills, as evidenced by their grade of 80 in CM001, a foundation course in academic and professional communication. CM002 builds upon this foundation by focusing on advanced skills in academic and professional communication, including advanced skills for searching academic databases, evaluating sources and incorporating and citing them appropriately in writing, as well as improving written and oral workplace communication. Therefore, CM002 is a suitable next step for this student, allowing them to further develop their already strong communication skills.
C	Based on the student's transcript, they have successfully completed CM001: a foundation course in communication for academic and professional contexts. They achieved a grade of 80, indicating a good understanding of the foundational communication skills. CM002, the recommended course, is an advanced course building upon these foundational skills. It focuses on advanced academic literacy, information literacy, and professional communication skills such as writing emails, meeting participation, client interaction, and professional image projection. Given the student's prior success in CM001 and their demonstrated proficiency in communication-related skills, CM002 appears to be a suitable next step in their academic journey. The advanced nature of CM002 aligns perfectly with the student's existing knowledge and ability, offering an opportunity to further refine and enhance their communication capabilities for both academic and professional success.
D	Based on the student's academic record, the recommended course CM002 is suitable. The student has demonstrated strong writing and presentation skills in CM001, achieving a grade of 80. This course covered various writing formats, including proposals, reports, and presentations, directly aligning with the skills needed for CM002's focus on FYP reports and technical presentations. The high grade suggests a solid foundation in these areas. Additionally, the student's strong performance in courses like Data Science and AI and Algorithm Design and Data Structures, which required technical report writing and presentations, further supports their readiness for CM002. The student's successful completion of multiple engineering courses also implies a capacity for independent research, a vital aspect of the FYP project.

LLM to respond according to a specific predefined perspective. By defining a role (e.g., a university course advisor), its output emphasizes salient aspects of the explanation making it relatable, contextual, and relevant to the learner. Cognitive verifier pattern, on the other hand, directs the LLM through a structured reasoning process. More specifically, by decomposing complex queries into simpler sub-questions or verification steps, the model focuses on specific details and exploits domain-specific knowledge. This process not only ensures accuracy and contextual relevance but also adherence to a coherent logical structure [26].

2. **Types of course descriptions:** Two broad methods are commonly adopted for drafting course descriptions: *Outcome-Based Course Description:* This approach is founded on the principles of Outcome-Based Education (OBE), which focuses on defining specific learning outcomes that students are expected to attain [8]. These outcomes form the basis for developing the skills and knowledge required for learners. In our data set, Outcome-Based Course Descriptions are drafted by course coordinators and contain course objectives and a set of intended learning outcomes. *Content-Based Course Descriptions:* This approach is centered around the subject matter and course content. It incorporates a comprehensive overview of the topics covered in chronological order and the learning methods employed, ensuring that students have a good understanding of the academic content to which they will be exposed [6].While this type of description helps students make informed decisions based on their interests and academic needs, course descriptions, taken in isolation, do not establish relationships between the courses.

3. **Social science theory:** Hilton's Conversational Model of Explanation posits that the primary objective of an explanation is to bridge knowledge gaps. By viewing explanation as a process where the explainer actively connects core issues with the broader context, the theory highlights the importance of relevance and clarity in an explanation [9]. For our context of course recommendation systems, this theory directs the creation of explanations that justify the recommendation of specific courses and provides the rationale in terms of relevance to the student's past academic achievements such as courses taken and their associated grades. Hence, the explanation becomes a tailored narrative that addresses individual knowledge gaps and fosters a deeper understanding of the recommendation process.

To evaluate the effectiveness of different explanation strategies, we compared four models that adopt the same meta-prompt described in Figure 2. These models differ only in their specific variations in prompt patterns, course description types, and the incorporation of social science theories. These variations are detailed in Table 1. Given the eight possible combinations that exist across the three dimensions,

Table 3: Student Selection

Student Type	Student Profile
Top Performers	Two students who consistently achieve high grades across all courses.
Average Performers	Two students with average performance in most courses.
Struggling Performers	Two students who have encountered difficulties in coursework, including failing some courses.

models A-D have been selected as the representative models for evaluation. In particular, given the computational constraints and the human expert workload required for evaluation, these four models provide a balanced and sufficient basis to draw initial conclusions about the impact of each design choice. Examples of explanations generated by four different models are shown in Table 2. These examples correspond to the recommendation of a communication course CM002.

Each pipeline model was implemented using the Gemini 1.5 Flash LLM model to generate explanations. Comparing these pipeline models allows us to assess how different prompting strategies and theoretical frameworks influence the clarity, relevance, and usefulness of the generated explanations.

2.3 Data collection and evaluation method

We evaluate the quality of explanations generated by different models by conducting a study involving university students, faculty members, and educational data mining researchers.

While the data set consists of 2,795 engineering students for training the recommendation model, a total of six students with diverse academic backgrounds were selected for the evaluation study. The students were grouped into the following categories based on their academic profile in Table 3. For each student, five courses were selected based on the recommendations from the deep learning model. This yielded a total of 120 explanations that were evaluated. This selection allows the evaluation of the framework's effectiveness across various academic profiles.

An exhaustive analysis was conducted using the explanations provided to a select group of students. Course instructors were invited to evaluate these explanations due to their role in crafting some of the course descriptions. They also serve as suitable assessors for the generated explanations. We modeled three distinct scenarios likely to occur in a real-world student course selection context:

1. Stakeholders, such as student care managers, who possess a comprehensive overview of course information.
2. Senior course instructors with substantial experience.
3. Junior course instructors with knowledge limited to a few related courses.

To provide a comprehensive overview of the available perspectives, three course instructors from the same academic institution as the students were involved. The expert panel consisted of an experienced course coordinator, with extensive experience in teaching engineering students and in-depth understanding of the curriculum, a senior course instructor, and a junior course instructor. Each expert independently evaluated the explanations' quality and effectiveness. To ensure a comprehensive evaluation, four scoring criteria [3] were established: Clarity, Effectiveness, Relevance, and Specificity, as detailed in Table 4:

Table 4: Four scoring criteria

Criterion	Explanation
Clarity	The explanation is easy to comprehend, with clearly stated reasoning.
Effectiveness	The explanation provides meaningful insights into the student's academic profile and aids decision-making for course selection.
Relevance	The explanation establishes accurate causal reasoning based on course descriptions and prior student performance.
Specificity	The explanation is sufficiently detailed and personalized, avoiding generic statements that could apply to a wide range of students.

2.4 Instrument: survey and open-end questions

The survey for human evaluators consists of two main sections: evaluation of explanations and open-ended feedback. For the former, a set of recommendations was presented to experts, accompanied by explanatory statements. The experts were invited to rate each explanation based on clarity, effectiveness, relevance, and specificity along a six-point Likert scale, with 1 representing "strongly disagree" and 6 representing "strongly agree." In addition to the explanatory statements, the student and courses profiles were provided for context.

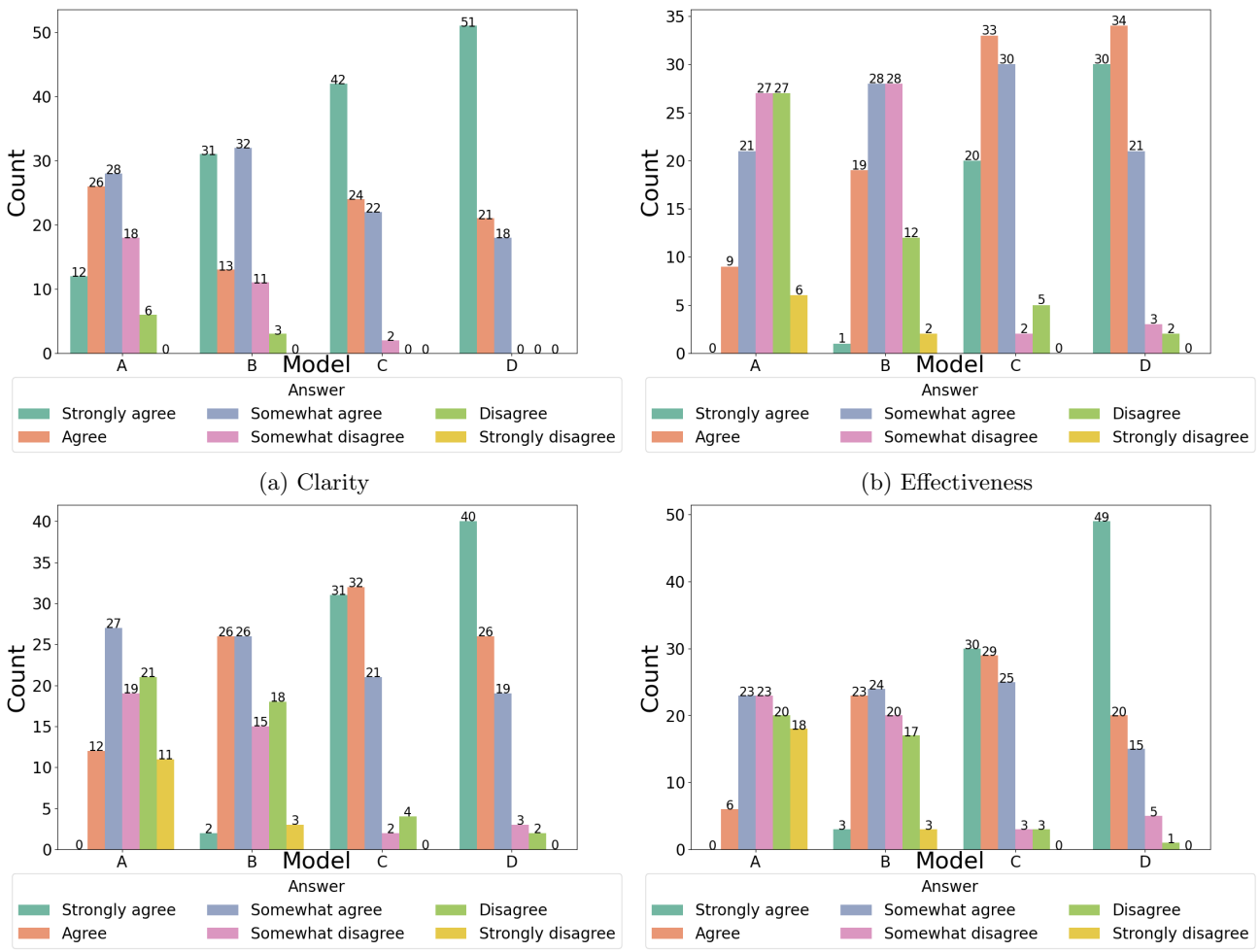
Post-Rating Reflections: Following the Likert-scale evaluations, the experts were invited to provide additional reflections through a set of multiple-choice and open-ended questions. These questions aimed to gather broader impressions of the explanations' effectiveness in supporting students, as well as suggestions for potential improvements. The survey was administered via Microsoft Forms, with participants receiving links containing detailed instructions and the full set of survey items.

3. RESULTS EVALUATION

3.1 Overall performance

Inter-rater reliability was computed using the Intraclass Correlation Coefficient (ICC) across the three experts. In particular, ICC(3,1) was used, which estimates reliability based on a single-rating, absolute-agreement, two-way mixed-effects model. ICC values for clarity, effectiveness, relevance, and specificity were 0.216, 0.437, 0.481, and 0.590, respectively. Experts also showed borderline moderate to moderate agreement on effectiveness, relevance, and specificity, using a specialized category approach; however, the lower clarity score indicates more subjectivity related to this criterion, possibly due to different expertise. This analysis indicates which expert can benefit the most from the explainable course recommendation system.

As shown in Figure 3, the performance of Model A through



(a) Clarity

(b) Effectiveness

(c) Relevance

(d) Specificity

Figure 3: Overall Performance of the Four Explanation Models on Four Scoring Criteria.

D is assessed across clarity, effectiveness, relevance, and specificity. Each criterion is assessed by three evaluators, who rated 30 explanations per model, yielding 90 ratings per model and 120 per evaluator per criterion. The ensuing discussion will elaborate on the findings and their implications.

Models C and D dominate in clarity, with 73%(66/90) and 80%(72/90) of their explanations rated as clear ("Strongly agree" or "Agree"), respectively. Model D achieved universal moderate-to-high clarity (without "Strongly disagree" ratings), while Model A and B achieved a lower score at 42%(38/90) and 49%(44/90), respectively. This difference in score highlights how structured prompting (Cognitive Verifier prompting pattern in Model C/D) and relevance theory enhance comprehensibility—having more structured and detailed prompt patterns significantly improve the comprehensibility of the explanations. Incorporating relevance theory appears to improve clarity by guiding the model to generate more context-aware explanations. Model A and B likely suffer from vague or generic outputs because of simpler prompting strategies.

Effectiveness measures the degree to which the explanations help students make informed course selections. The result shows that Model C and D outperform A and B. Specifically, Model D achieved 71%(64/90) of its explanations as "Highly effective" as opposed to 59%(53/90) for Model C. In contrast, Models A and B scored a modest 50%. This suggests that content-based explanations (e.g., detailing course topics, skills, or prerequisites) are more actionable for students than outcome-based descriptions (e.g., "improve critical thinking"). Incorporating detailed course content has a more significant impact on students' decision-making processes than merely presenting learning outcomes. These results imply that explanations that focus on content help students better align course recommendations with their academic goals.

Relevance, which evaluates connections between past and recommended courses, shows Model D excelling: 44%(40/90) of its explanations received "Strongly agree" ratings, and 94% achieved at least moderate relevance. A clear performance gradient (A < B < C < D) aligns with increasing prompt complexity. For example, Model D's instructions to "identify shared skills, prerequisites, or thematic

Table 5: Average rating for different criterion on different model.

	Clarity	Effectiveness	Relevance	Specificness
A	4.2	3	3.1	2.8
B	4.6	3.6	3.7	3.6
C	5.2	4.7	4.9	4.9
D	5.4	5.0	5.1	5.2

links" likely enabled richer relational reasoning. Content-based approaches (D) also outperformed outcome-focused ones (A/B/C), as topics and skills are more concrete than abstract outcomes for establishing relevance.

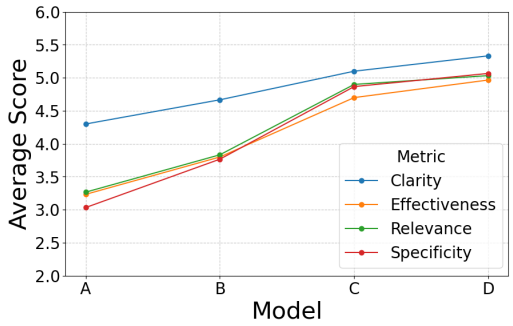
Model D achieves the best performance in terms of specificity, with 54%(49/90) of its explanations rated "Strongly agree," compared to Model C's 67%(60/90) in the top two categories. This highlights its capability to generate highly personalized and detailed explanations tailored to individual student profiles. Model A/B received predominantly neutral/disagree ratings. This modest performance suggests that they fail to provide adequately detailed or personalized recommendations for students with varying academic backgrounds. With more detailed prompting patterns provided, the explanations tend to be more personalized.

The evaluation reveals consistent superiority of Model C and D over Model A and B across all criteria, as summarized in Table 5. Model C and D achieved average ratings 1.5–2 times higher than their counterparts, underscoring the transformative impact of structured prompt design and content-driven explanations. The performance gap stems from three interrelated factors. First, Cognitive Verifier prompting pattern enhanced clarity and specificity by enforcing logical flow and tailoring outputs to individual academic profiles. Second, content-based explanations, which emphasize course topics, exhibited more actionable and relevance than outcome-focused descriptions. By grounding recommendations in tangible academic elements, Model C and D helped students map suggestions to their goals, whereas vague outcomes in Model A/B led to generic or unconvincing explanations. Third, the integration of relevance theory in Models C and D, which prioritizes contextual relationships such as shared skills or thematic overlap between courses, amplified their ability to establish meaningful connections. This approach not only improved relevance ratings, but also strengthened perceived effectiveness, as students could better trace the rationale behind recommendations.

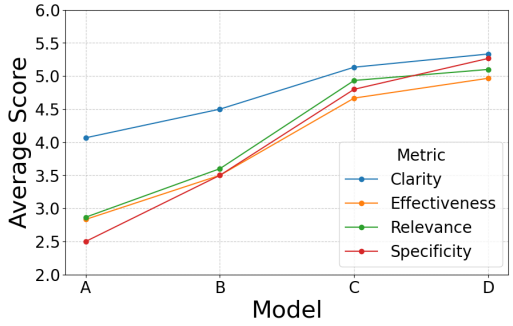
3.2 Model performance across student academic profiles

Figure 4 illustrates the performance of four models (A, B, C, and D) across four evaluation metrics for students at various academic performance levels: top performers, average performers, and struggling performers.

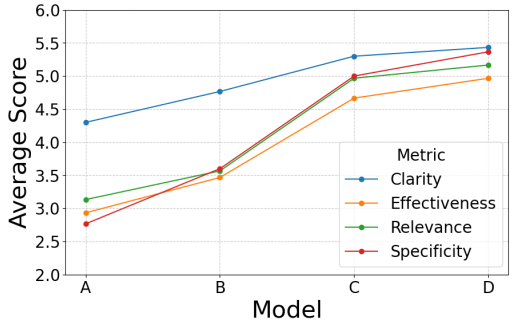
A consistent trend of increasing performance from Model A to Model D is observed across all student categories when evaluated against the four scoring criteria. Notably, Model D consistently outperforms the other models across the student groups, regardless of academic standing. Top Performers: Model C and D deliver high-quality explanations, with average scores exceeding 4.0 across all criteria. Average Performers: Model D demonstrates consistent performance across all four criteria, with scores consistently above 4.2. Model C also maintains competitive scores, ranging from 4.0 to 4.3 across all metrics, showing its adaptability to students with balanced academic profiles. Struggling Performers: Model C significantly outperforms Model A and B, with scores nearing 4.0 in all four categories. Furthermore, an improvement in performance by Model D is evident, with an average increase of 0.5 to 0.7 points, particularly in Effectiveness and Specificity. This finding suggests that Model D has the ability to provide tailored support for underperforming students.



(a) Top Performers.



(b) Average Performers.



(c) Struggling Performers.

Figure 4: Performance of the Four Models on Different Student Category.

The consistent and comparable performance of Model C and D across all metrics and student groups suggests strong generalizability, effectively serving students with diverse aca-

demic profiles without bias. Model D, in particular, demonstrates robust performance, maintaining high scores in all student categories, indicating a strong capacity to provide tailored high-quality recommendations that align with individual academic needs. Neither model exhibits a significant decrease in scores for any student group, further supporting their efficacy across varied academic profiles.

3.3 Rating trends across different experts

As illustrated in Figure 5, the ratings provided by three different experts—experienced course coordinator, senior course instructor, and junior course instructor—are presented for the four models (A, B, C, and D) in four evaluation criteria. It is evident that the three expert groups exhibit a consistent consensus that the performance of the model undergoes a progressive enhancement from Model A to Model D in all four criteria, which is in line with the conclusion from Figure 3 and Figure 4. It is useful to note the strong consensus among the three expert groups on the superior performance of Model C and D, which exhibited a remarkable capacity to generate clear, effective, and pertinent explanations.

In addition, we also note that experienced course coordinators with a deep understanding of course structures and academic requirements, impose a more rigorous evaluation standard. This is supported by their lower ratings compared to the senior and junior instructors, implying a modest level of support by the generated explanations to learners with extensive knowledge of the course settings. In contrast, senior and junior course instructors provide analogous ratings across the majority of criteria, particularly with regard to Model C and Model D. This congruence indicates that both groups find the explanations generated by these models to be of significant value in facilitating the course selection process. These explanations likely offer actionable insight and clear reasoning to support instructional needs and teaching responsibilities such as academic advisory tasks.

Based on the above findings, we conclude that the explanations generated are more instructive and effective for stakeholders with less experience in course design and instruction. More specifically, the detailed analyzes and structured recommendations generated by Model C and D appear to be beneficial to these groups—they can potentially support decision making and improve understanding of the relevance of the course to the academic profile and prior courses. Furthermore, we infer that students who, in general, are less familiar with course structures and content compared to instructors, are likely to find these explanations helpful. This finding underscores the significant potential of models such as C and D in helping the students understand how the recommended (candidate) courses may scaffold prior courses to support the course selection process. These models can, therefore, provide customized guidance and support meaningful academic decision making, contributing thus to the holistic development of students.

3.4 Insights from open-ended responses

The human evaluators were also invited to offer qualitative feedback regarding the aspects that rendered the explanations beneficial and the aspects that can further be improved. The feedback provided by the experts varied in its perspective, and the evaluation of the explanations exhibited significant variation depending on the model employed. Among the models that were considered beneficial, experts highlighted the importance of presenting the grades of previously undertaken courses and how these courses are associated with the content of the recommended courses to assess the quality of the explanation.

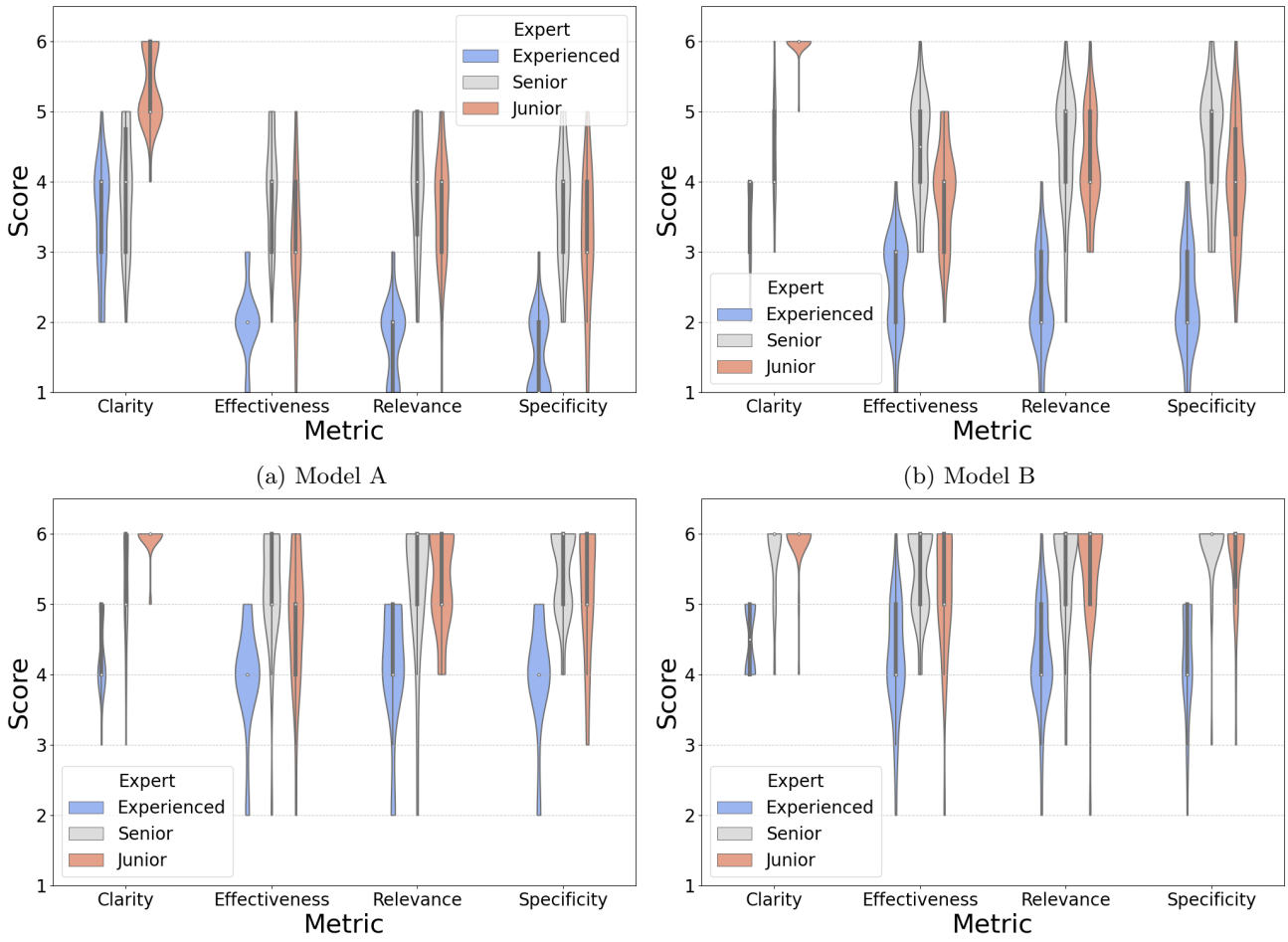
To improve the clarity of the explanations, the evaluators suggested incorporating visualizations to highlight the relationships between courses, thus improving comprehension. They also highlight the usefulness of incorporating quantitative information (e.g., grade point average associated with previous courses or predicted grades for the recommended courses) to understand the rationale behind the generated recommendations. Although achieving high grades may not be the only factor influencing course selection, providing "explanation behind the explanation" offers greater insights and improves trust in the recommendation systems. In summary, the quality of the explanations is model-dependent—Model C and D were deemed adequate in generating helpful explanations as these models provided sufficient information on prior performances and clearly established the correlations with the recommended courses. Extending beyond the use of text by incorporating a variety of presentation styles will enhance the comprehensibility of the generated explanations. Moreover, the integration of information derived from predictive models or other educational models has been identified as a potential strategy to augment the effectiveness of the explanations, a benefit that would be further compounded if these models were to be integrated with existing recommendation systems.

4. DISCUSSION

The primary contributions of this study are as follows:

1. The development of an LLM-based course description-driven framework to generate natural language explanations for AI course recommendation systems;
2. The evaluation of the generated explanations based on academic profiles and comparison of different prompting patterns;
3. The analysis of the effectiveness and key factors contributing to good explanations.

In conclusion, this study offers an important contribution to LLM-based explanable course recommendations driven by course descriptions, specifically Model C and D, as a support for the student course selection process. Regarding the first research question, the effectiveness of the explanation is model-dependent—Model C and Model D achieved satisfactory results. Moreover, comparing the ratings by different experts indicates the potential of acceptance and effectiveness for students who lack expertise and experience in the courses and curriculum. As for the key factors that can help generate effective explanations, based on the analysis of the rating results, a more detailed prompt pattern would benefit clearer text generation. By translating the theory of relevance [9], the model would be better equipped to explore the relevance among courses and explain them in detailed topics, which in turn makes the explanations more helpful and personalized. By utilizing content-based course descriptions, more effective and personalized explanations can be achieved.



(a) Model A

(b) Model B

(c) Model C

(d) Model D

Figure 5: Rating of the Four Models on Different Experts.

The satisfactory results for both Model C and Model D indicate that the different types of course descriptions used may contribute to the effectiveness of explanation generation. The study could be generalized to students with different academic abilities by providing customized instructions and to various learning contexts, including K-12 education and online learning environments such as intelligent tutoring systems and Massive Open Online Courses (MOOCs) [10], provided the learning context is fully specified. The objective of this work is to enhance the interpretability of course recommendation systems, with the aim of improving the educational experience of students and supporting educators in making data-driven decisions.

5. LIMITATIONS AND FUTURE WORK

The study has several limitations. First, the evaluation was performed from the course coordinators' and educators' perspective. To develop a more holistic understanding, further evaluations should include other stakeholders in the processes around developing student course preferences (e.g., students, student care managers, academic administrators, and mentors). Their insights can offer a broader range of viewpoints. In addition, the current study involved a limited number of experts. Future research focusing on a wider population and more experts would yield more robust and generalizable findings.

The present study focuses on explanations based solely on course descriptions and presented in text format. In subsequent research efforts, the incorporation of additional modalities, including visual aids and interactive elements, has the potential to improve the clarity and efficacy of the explanations. This multimodal approach has the potential to better communicate information to various stakeholders. Moreover, the integration of the explainable course recommendation system with other educational applications could facilitate more detailed analysis and support student course selection. This integration could offer a more comprehensive educational decision support tool.

6. ACKNOWLEDGMENTS

The authors would like to thank Dr. S Supraja who contributes her time and expertise to evaluate the explanable course recommender's performance.

References

[1] Hasan Abu-Rasheed et al. "Supporting student decisions on learning recommendations: An llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring". In: *arXiv preprint arXiv:2401.08517* (2024).

[2] Muhammad Afzaal et al. "Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation". In: *Frontiers in Artificial Intelligence* 4 (2021), p. 723447.

[3] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey". In: *arXiv preprint arXiv:2006.14799* (2020).

[4] Cristina Conati et al. "Toward personalized XAI: A case study in intelligent tutoring systems". In: *Artificial intelligence* 298 (2021), p. 103503.

[5] Jacob Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[6] Robert M Diamond. *Designing and assessing courses and curricula: A practical guide*. John Wiley & Sons, 2008.

[7] Gökhan Engin et al. "Rule-based expert systems for supporting university students". In: *Procedia Computer Science* 31 (2014), pp. 22–31.

[8] Ronald M Harden. *Developments in outcome-based education*. 2002.

[9] Denis J Hilton. "A conversational model of causal explanation". In: *European review of social psychology* 2.1 (1991), pp. 51–81.

[10] Andreas M Kaplan and Michael Haenlein. "Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster". In: *Business horizons* 59.4 (2016), pp. 441–450.

[11] Md Akib Zabed Khan, Agoritsa Polyzou, et al. "Session-based Methods for Course Recommendation". In: *Journal of Educational Data Mining* 16.1 (2024), pp. 164–196.

[12] Hassan Khosravi et al. "Explainable artificial intelligence in education". In: *Computers and Education: Artificial Intelligence* 3 (2022), p. 100074.

[13] Jiawei Li et al. "Grade Prediction via Prior Grades and Text Mining on Course Descriptions: Course Outlines and Intended Learning Outcomes." In: *International Educational Data Mining Society* (2022).

[14] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81.

[15] Boxuan Ma, Yuta Taniguchi, and Shin'ichi Konomi. "Course Recommendation for University Environments." In: *International educational data mining society* (2020).

[16] Boxuan Ma, Tianyuan Yang, and Baofeng Ren. "A Survey on Explainable Course Recommendation Systems". In: *International Conference on Human-Computer Interaction*. Springer. 2024, pp. 273–287.

[17] Kishore Papineni et al. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[18] Zachary A Pardos and Weijie Jiang. "Designing for serendipity in a university course recommendation system". In: *Proceedings of the tenth international conference on learning analytics & knowledge*. 2020, pp. 350–359.

[19] Pat Pataranutaporn et al. "AI-generated characters for supporting personalized learning and well-being". In: *Nature Machine Intelligence* 3.12 (2021), pp. 1013–1022.

[20] Muh Putra Pratama, Rigel Sampelolo, and Hans Lura. "Revolutionizing education: harnessing the power of artificial intelligence for personalized learning". In: *Klasikal: Journal of education, language teaching and science* 5.2 (2023), pp. 350–357.

[21] W Qiu et al. ""I Am Here To Guide You": A Detailed Examination of Late 2023 Gen-AI Tutors Capabilities in Stepwise Tutoring in an Undergraduate Statistics Course". In: *INTED2024 Proceedings*. IATED. 2024, pp. 3761–3770.

[22] Wei Qiu, S Supraja, and Andy W. H. Khong. "Toward Better Grade Prediction via A2GP-An Academic Achievement Inspired Predictive Model". In: *Proceedings of the 15th International Conference on Educational Data Mining*. 2022, p. 195.

[23] Mehbooba Shareef et al. "Advancing explainable MOOC recommendation systems: a morphological operations-based framework on partially ordered neutrosophic fuzzy hypergraphs". In: *Artificial Intelligence Review* 58.2 (2024), p. 46.

[24] Vinitra Swamy et al. "iLLuMinaTE: An LLM-XAI Framework Leveraging Social Science Explanation Theories Towards Actionable Student Performance Feedback". In: (2025). Accepted at AAAI 2025. URL: `https://arxiv.org/abs/2409.08027`.

[25] Thomas Yu Chow Tam et al. "A framework for human evaluation of large language models in healthcare derived from literature review". In: *NPJ digital medicine* 7.1 (2024), p. 258.

[26] Jules White et al. "A prompt pattern catalog to enhance prompt engineering with chatgpt". In: *arXiv preprint arXiv:2302.11382* (2023).

[27] Junchao Wu et al. "A survey on LLM-generated text detection: Necessity, methods, and future directions". In: *Computational Linguistics* (2025), pp. 1–66.

[28] Run Yu et al. "Orienting students to course recommendations using three types of explanation". In: *Adjunct proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 2021, pp. 238–245.