From Frustration to Personalization: Revolutionizing Course Recommendations with LLMs

Ruitao Lai
University of Toronto
Toronto, Ontario, Canada
r.lai@mail.utoronto.ca

Yuzhi Tang
University of Toronto
Toronto, Ontario, Canada
yuzhi.tang@mail.utoronto.ca

ABSTRACT

The process of course selection is pivotal for students, shaping their academic performance, career trajectories, and personal growth. However, students often face challenges such as aligning course choices with goals, managing workloads, and navigating institutional requirements. Traditional recommendation systems, based on structured data, fail to capture the nuances of individual preferences. This study addresses these limitations by introducing a Large Language Model (LLM) framework integrated with Retrieval-Augmented Generation (RAG). By analyzing real-world student discussions and leveraging data from the University of Toronto, the research identifies five key challenges in course selection: relevance, workload management, scheduling conflicts, program requirements, and teaching quality. The proposed LLM-powered system dynamically processes natural language queries, providing personalized, transparent, and relevant recommendations. Evaluation results demonstrate its strength in relevance, personalization, and transparency, with room for improvement in fostering academic exploration. This work represents a step toward revolutionizing course recommendation systems to better meet the multifaceted needs of students.

1 INTRODUCTION

Course selection is a critical, high-stakes decision-making process that profoundly influences students' academic progress, graduation outcomes, and long-term career trajectories [2, 7, 9]. Students face a wide range of considerations when selecting courses, including ensuring alignment with program requirements, managing course load to optimize performance, and identifying classes that match their academic interests and professional aspirations [2, 5]. External factors, such as balancing work and family commitments, also shape how students approach this task, especially in diverse educational contexts [9]. As higher education institutions expand their program offerings and online enrollment opportunities, the complexity of choosing a suitable combination of courses continues to grow, increasing the need for personalized and adaptable guidance.

Traditional course recommendation systems have leveraged collaborative filtering, content-based filtering, or hybrid approaches to match students with suitable courses [6, 10]. While these methods offer valuable insights, they often rely on structured data or explicit student feedback, limiting their ability to fully capture the nuances of individual preferences and the evolving academic landscape. In recent years, the emergence of large language models (LLMs) has opened new opportunities to enhance recommendation systems by understanding complex, natural language queries and reasoning about user needs within richer textual contexts [3, 4]. LLMs, such as GPT-based models, have demonstrated remarkable capabilities in encoding semantic meaning and generating coherent, human-like responses, making them well-suited for educational applications that involve interpreting user goals and course content in a more flexible manner.

Against this backdrop, the integration of LLMs into course recommendation workflows represents a promising direction for improving personalized academic guidance. By leveraging large-scale pretrained language models, these systems can interpret intricate user queries, dynamically adapt recommendations based on shifting educational requirements, and provide transparent, human-readable justifications for their suggestions. Recent work in the recommendation literature has begun exploring the use of LLMs to enhance semantic understanding, incorporate external knowledge sources, and generate explanations that increase user trust [8, 11]. In the context of course selection, an LLM-driven recommender can potentially alleviate common challenges—such as identifying courses that satisfy program prerequisites, balancing course load while maintaining academic performance, or encouraging exploration beyond a student's initial field of interest [1, 5, 7].

This paper examines how LLM-based methods can address the multifaceted challenges students face when selecting courses, particularly at large and diverse institutions like the University of Toronto. We present a Retrieval-Augmented Generation (RAG) framework that integrates LLMs with a vector database of course information, enabling flexible, text-based query processing and improved alignment with users' individualized preferences. Our approach builds on recent advancements in LLM-driven recommendation and natural language understanding, contributing both a practical system for course selection and a validation methodology that leverages automated LLM-based evaluators, supplemented by human judgment for credibility checks.

2 RELATED WORK

Course Selection. Course selection is a crucial process that significantly influences students' academic journeys and future career trajectories [7]. The number of courses students choose to enroll in each semester, known as their course load, along with the specific courses they select, can have a substantial impact on their academic performance [2, 9]. A heavier course load can lead to lower academic performance, while a strategically managed course load, especially in the initial stages of university, has been linked to positive academic results, such as higher GPAs and a higher probability of graduating on time [2, 9]. This association between course load and academic performance is further emphasized by findings that a higher course load in the second semester is positively associated with students' academic performance over four

years, while a higher course load in the fifth and sixth semesters is negatively associated with performance [7].

Several factors, both environmental and psychological, contribute to students' course load decisions. In the US, the heterogeneity of the student population in terms of age and socioeconomic backgrounds results in varying work and family obligations, as well as financial burdens due to tuition fees, all of which can affect course selection and academic performance [2, 9]. Psychological factors, such as motivation, planning, and cognitive considerations, also play a role [7]. Students with stronger motivation tend to procrastinate less and exhibit proactive learning behaviors [7]. Effective planning, a key aspect of metacognition and intelligence, is particularly evident among students who perceive university as a pathway to their desired careers and strategically plan their coursework to align with future internships or research opportunities [7]. Additionally, cognitive considerations can influence course load decisions, with students experiencing lower academic performance potentially opting for fewer courses initially to adjust to university life and manage their workload effectively [7].

Beyond course load, the specific courses students choose also reveal connections to their academic performance and psychological processes [5]. A diverse course selection, measured by the number of different departments from which students take courses, is associated with work mastery, competitiveness, and academic performance [5]. The tendency for students with lower academic performance to shy away from demanding courses suggests a strategic approach to course selection [1]. Furthermore, academic performance can be linked to specific course choices, implying that intrinsic and extrinsic motivations and cognitive ability influence these decisions beyond mere interest in course content or anticipated classroom experiences [5]. These findings emphasize the need to account for potential biases in measuring academic performance, particularly grade leniency, which can stem from variations in grading standards across courses and instructors [1].

Overall, course selection is a multifaceted decision-making process shaped by a complex interplay of environmental and psychological factors. Students engage in this process while considering their academic abilities, motivations, and future aspirations. Recognizing the significance of these factors and providing students with personalized guidance and support, such as effective academic advising, is crucial in empowering them to make informed choices that align with their academic goals and career ambitions [7, 9].

LLMs for Recommendation System. Recent surveys have highlighted the potential of LLMs to tackle longstanding challenges in recommendation research, including the "cold start" problem and the need for more transparent, explainable recommendations. Li et al. (2024) provide a comprehensive overview of how LLMs can enrich recommendation systems by improving semantic understanding, user intent interpretation, and cross-domain generalization [8]. Similarly, Wu et al. (2024) discuss the prospects and emerging frontiers of LLM-based recommendation, noting their ability to incorporate unstructured textual data and provide nuanced rationales for recommended items [11].

These advancements align closely with the goal of enhancing course recommendation systems. Traditional course recommenders have relied on collaborative filtering, content-based approaches, or hybrid methods to match students with relevant courses based on historical enrollment patterns, ratings, or degree requirements [6]. While these techniques are effective at surfacing relevant courses, they often lack the flexibility to process complex, open-ended queries or generate human-readable explanations for their suggestions. By contrast, LLM-based approaches can parse natural language queries from students, understand their academic interests and constraints, and produce rich, narrative-style justifications—facilitating more personalized and transparent advising experiences.

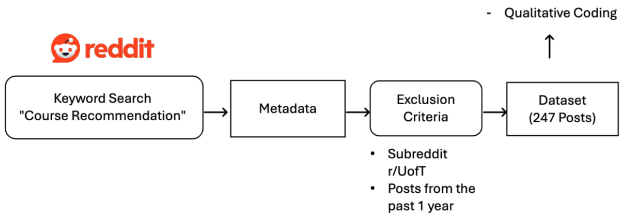

Figure 1: Overview of the data collection process for RQ1.

3 Data Collection

In this section, we describe the dataset employed in our study (Section 3.1), outline the preprocessing methods applied to the dataset (Section 3.2). Figure 1 provides an overview of the data collection process and highlights the datasets utilized for research question 1.

3.1 Data Source

To collect our dataset, we scraped posts from Reddit, an online forum frequented by University of Toronto students to discuss academic topics, including course recommendations. The data collection process involved extracting both the titles and bodies of posts to capture the full context of each discussion. We targeted posts explicitly related to course selection by filtering for keyword "course recommendation." This keyword were chosen to ensure that the collected posts were relevant to the topic of course recommendations.

The subreddit was selected for its active participation and diverse user base, which provides authentic insights into student concerns and experiences. The scraping process adhered to ethical guidelines, ensuring that only publicly available posts were collected, and no personally identifiable information was included in the dataset. By focusing on a student-driven platform, we aimed to build a dataset that reflects real-world discussions and challenges faced by students in selecting courses. This data collection approach ensures that the dataset is both comprehensive and representative of the experiences shared by students in this academic community.

3.2 Data Preprocessing

In our analysis, we focus exclusively on Reddit posts related to course recommendations from the r/UofT subreddit. To ensure relevance and quality, we applied a series of data preprocessing steps. First, we restricted the dataset to include only posts from the r/UofT subreddit, excluding any posts from other subreddits, to ensure the

discussions captured were specific to the University of Toronto context. Second, we limited the dataset to posts created within the past year, excluding older posts to maintain the relevance of the data to current course selection practices. Third, we focused solely on the titles and main body text of posts, discarding comments and other metadata to concentrate on the original intent and context of the discussions. After applying these steps, we identified a total of 247 posts that met our inclusion criteria, forming the basis for our analysis of student challenges in course recommendations.

4 RQ1: What challenges do students face when receiving course recommendations?

4.1 Motivation

Understanding the challenges students face in receiving course recommendations is crucial to improving the effectiveness of recommendation systems. Students encounter a diverse range of obstacles, including difficulties in identifying courses that align with their academic goals, personal interests, or career aspirations. Logistical issues, such as scheduling conflicts and the lack of remote or online course options, further complicate the selection process. Additionally, many students seek courses that balance manageable workloads with the potential for high grades, often expressed as a preference for "bird courses." Institutional requirements, such as fulfilling program prerequisites or graduation criteria, add another layer of complexity. Lastly, concerns about teaching quality and professor effectiveness highlight the qualitative aspects students consider when selecting courses. By exploring these multifaceted challenges, we aim to uncover patterns and gaps in existing recommendation systems, providing insights that can guide the development of more tailored and effective solutions.

4.2 Approach

We applied an open coding process, involving two independent coders and multiple rounds of refinement. Our labeling process was conducted over three stages:

(1) **Initial Coding and Codebook Formation:** Both coders independently analyzed a random subset of 30 posts from the dataset to identify preliminary themes. Following this, they engaged in a discussion to compare their findings and collaboratively develop a coding book that provided clear definitions and criteria for each emerging code.
(2) **Refinement and Inter-Rater Reliability Assessment:** Using the established coding book, both coders independently applied the codes to another random subset of 30 posts. The inter-rater agreement was measured using Cohen's Kappa coefficient, achieving a score of 0.74, which indicates substantial agreement. Discrepancies were discussed, and the coding book was further refined for clarity and consistency.
(3) **Final Application to the Full Dataset:** The finalized coding book was applied independently by both coders to the full dataset. Any remaining disagreements were resolved through discussion, ensuring that the final taxonomy was comprehensive and representative of the data.

Challenges	Count
Seeking specific course recommendations	88
Looking for easy courses (bird courses)	47
Looking for courses matching personal interests	13
Fulfilling program or graduation requirements	12
Dealing with scheduling conflicts	9
Overwhelmed by choices or workload	8
Preferring online or remote courses	5
Seeking courses aligned with career goals	5
Seeking advanced or specialized courses	3
Looking for beginner-level courses	2
Focused on maximizing GPA or grades	2
Concerned about professor or teaching quality	2
Unclear	49

Table 1: Taxonomy of the initial identified challenges and their corresponding counts.

4.3 Results

Table 1 presents the taxonomy of challenges students face in course recommendation systems, derived from analyzing 247 posts on the r/UofT subreddit. These challenges have been categorized into five major themes (C1 to C5) that represent the primary concerns and intentions expressed by students in their course selection process.

(C1) Course Suitability and Relevance: This category encompasses the largest proportion of student posts, highlighting the importance of selecting courses that align with their academic or personal objectives. A significant number of posts specifically request recommendations for courses that match their interests, career goals, or academic requirements. Subcategories within this theme include students seeking advanced or specialized courses to deepen expertise in their field, beginner-level courses to build foundational knowledge, and general courses for personal enrichment. Posts in this category reflect a strong demand for courses that are perceived as both relevant and beneficial to individual aspirations.

(C2) Program and Graduation Requirements: Many students express a need for guidance in selecting courses that fulfill mandatory program prerequisites or graduation requirements. This category reflects the institutional constraints faced by students, where their course choices are often dictated by academic policies. Posts in this category typically request information about courses that meet specific credit or degree requirements, emphasizing a pragmatic approach to course selection.

(C3) Scheduling and Accessibility: Logistical challenges, such as scheduling conflicts and accessibility concerns, are prominent in this category. Students frequently highlight difficulties in fitting courses into their timetables, particularly when balancing other academic or personal commitments. Additionally, posts in this category include requests for online or remote courses, underscoring the growing demand for flexible learning options that accommodate diverse needs and circumstances.

(C4) Workload and Performance Concerns: A substantial number of posts reflect concerns about managing academic workloads while maintaining high grades. This category includes students explicitly searching for "bird courses," which are perceived as easy or less time-consuming, to balance their overall workload.

Others focus on finding courses that allow them to achieve high GPAs without excessive stress, highlighting the interplay between academic performance and course selection.

(C5) Instructor and Teaching Quality: Though less frequent, some students emphasize the importance of teaching quality and professor effectiveness in their course selection process. Posts in this category typically express concerns about unclear instruction, ineffective communication, or overall teaching methods. These discussions reflect a subset of students who prioritize learning experiences and pedagogical excellence when choosing courses.

Each category represents a distinct set of challenges faced by students in the course recommendation process, emphasizing the multifaceted nature of this decision-making task.

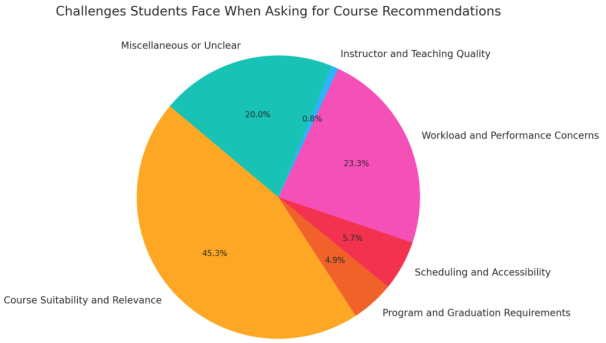

Figure 2: Pie Chart of Identified Challenges from r/Uoft Subreddit Posts

5 RQ2: How can LLM-based tools address the challenges in course recommendation systems?

5.1 Approach

We propose a Retrieval-Augmented Generation (RAG) framework (Figure 3) to build an LLM-powered course recommendation system that addresses the diverse challenges identified in RQ1. The framework integrates a vector database, query augmentation, and document retrieval to ensure that recommendations are grounded in curated course data while providing a user-centric experience.

Our implementation formulates course recommendation as a question-answering (QA) task, where users interact with the system by posing natural language queries about their academic needs. This approach mimics the interaction with a school counselor, enabling the system to provide personalized and flexible recommendations. To enhance retrieval accuracy, user queries are augmented by an LLM before being processed. Retrieved courses are ranked by the LLM based on their relevance to the user's preferences, and the LLM provides rationales for its recommendations. For all our experimentation, we used the GPT-4o-mini model as our LLM accessed via the OpenAI API.

5.1.1 Data Preparation. For this study, we used the University of Toronto - St. George Campus's course database, which includes 5,156 undergraduate courses across 118 programs. The course titles and descriptions were parsed from the Arts & Science course calendar (https://artsci.calendar.utoronto.ca/search-courses) on October 10, 2025. This dataset serves as the foundation for building the vector database. Although the data is publicly available, we implemented a RAG framework to mitigate the risks of hallucinations and outdated knowledge in pretrained LLMs.

5.1.2 Vector Database. The parsed course data was stored in a MongoDB database and converted into vector embeddings using the all-MiniLM-L6-v2 sentence transformer. The embeddings are 384-dimensional and they capture the semantic meaning of course descriptions, enabling efficient similarity-based retrieval. FAISS was selected to implement the vector database due to its scalability and performance in handling large-scale vector search tasks.

5.1.3 Query Augmentation. User queries are augmented to improve retrieval performance and better align with the semantic structure of course descriptions. The system passes the user's query to an LLM, which reformulates it by inferring intent and emphasizing key attributes such as subject area, workload, prerequisites, and scheduling preferences. For example:

- Original Query: "easy CS courses for beginners"
- Augmented Query: "introductory computer science courses requiring minimal prior knowledge and low workload."

The prompt passed to the LLM to reformulate the query is as follows:

> "You are helping a student find relevant courses. Rewrite the following query to better match the format and structure of course descriptions in a university catalog. Ensure that the rewritten query highlights the user's preferences, such as subject area, level of difficulty, prerequisites, scheduling needs, or workload expectations:"

This query rewriting step improves the likelihood of retrieving courses that match user preferences.

5.1.4 Document Retrieval. The augmented query is converted into a vector embedding using the same BERT model and compared against the course database to retrieve the top-k most semantically similar courses. For this study, k=10, balancing comprehensiveness with the LLM's context length limitations.

Courses retrieved from the vector database are passed to the LLM for ranking based on user-specific criteria such as relevance, workload, and alignment with academic goals. The ranking process is guided by the following prompt designed to ensure recommendations are personalized and contextual:

> "You are assisting a student in selecting courses. Based on the user's query and the following retrieved course descriptions, rank the courses from most (rank 1) to least (rank 10) relevant. Use the following criteria: relevance to the user's stated goals, workload alignment, and academic requirements. Provide a ranked list of the courses, assigning a rank (1 to 10) to each, but do not recommend any courses at this stage:"

5.1.5 Course Recommendation. The LLM generates final course recommendations based on the ranked list of retrieved courses and

Category	Challenges	Count
C1: Course Suitability and Relevance	Seeking specific course recommendations	88
	Looking for courses matching personal interests	13
	Seeking courses aligned with career goals	5
	Seeking advanced or specialized courses	3
	Looking for beginner-level courses	2
C2: Program and Graduation Requirements	Fulfilling program or graduation requirements	12
C3: Scheduling and Accessibility	Dealing with scheduling conflicts	9
	Preferring online or remote courses	5
C4: Workload and Performance Concerns	Overwhelmed by choices or workload	8
	Focused on maximizing GPA or grades	2
	Looking for easy courses (bird courses)	47
C5: Instructor and Teaching Quality	Concerned about professor or teaching quality	2
Unclear	Unclear or other challenges	49

Table 2: Summary of challenges and their counts categorized into six categories (C1–C5).

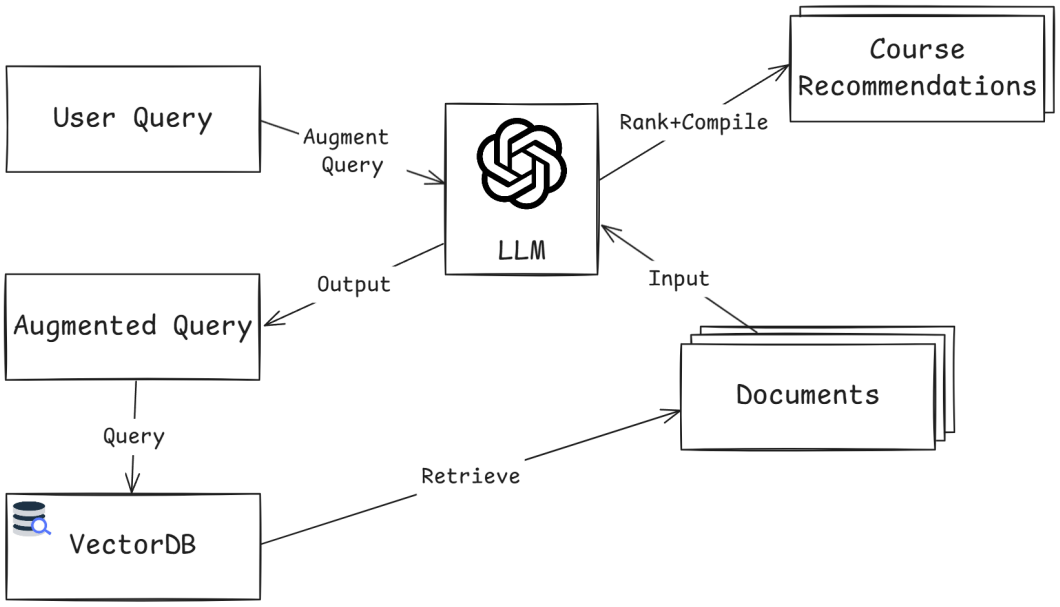

Figure 3: Overview of the LLM course recommender framework: A user query is first augmented by the LLM to enhance retrieval; the augmented query is used to retrieve the top-k course information from the vector database; the LLM ranks the courses based on relevance to the user query; finally, the LLM compiles the recommendations with detailed rationale.

the user query. A tailored prompt is used to guide this process, ensuring recommendations are drawn directly from the pre-ranked courses without re-ranking them:

> "Based on the user's query and the ranked list of courses below, recommend courses from the list that best match the user's preferences. Use the provided ranking to guide your selection. Provide a brief rationale for each recommended course, explaining why it aligns with the user's stated goals or needs. Do not alter the order or ranking of the courses:"

Rationales for each recommendation are generated by synthesizing attributes from the retrieved courses, such as subject area, prerequisites, and workload. While prerequisites are not explicitly considered in this study, they may be incorporated into future iterations of the system.

5.2 User Interface

We designed a web application with JavaScript and Node Package Manager (NPM) and a backend with Flask to facilitate course recommendations through an intuitive and user-friendly interface. The application opens with a welcome screen (Figure 4), and clicking the "Start" button scrolls the webpage to the dialogue interface (Figure 5). The interface emulates a standard QA chatbot, allowing users to interact with the system naturally. A trash icon enables users to clear the current conversation and explore multiple queries

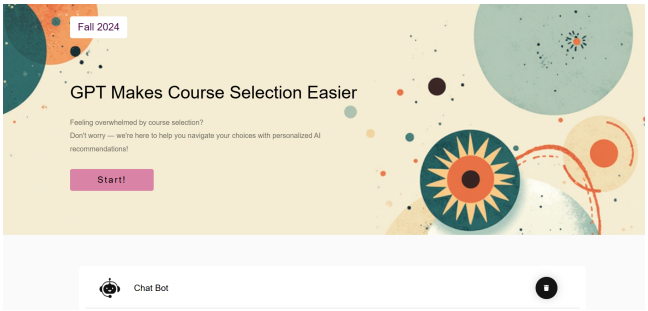

Figure 4: Welcome screen of the LLM Course Recommender website

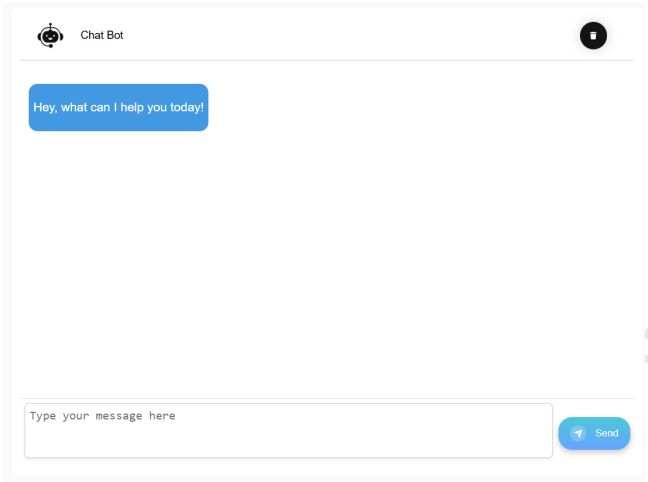

Figure 5: Dialogue screen of the LLM Course Recommender website

without refreshing the page. The UI is used in the human evaluation study to allow participants to interact with the system in real time and provide feedback.

5.3 Evaluation

The LLM course recommender needs to support queries from diverse student backgrounds and goals. Doing a full-scale human study would require recruiting participants from different programs of study at different levels, which is highly time and resource-intensive. We developed an automatic evaluation framework using LLMs to simulate students with diverse profiles and evaluate the quality of the course recommendations (Figure 6).

The framework includes testing the course recommender with query templates designed to span various capability categories (e.g. course suitability and relevance), and can be customized with different keywords to simulate diverse student goals, programs, and requirements. These templates are used to prompt the LLM recommender to generate course recommendations, which are then evaluated and scored by an LLM evaluator based on human-centered performance metrics. A small-scale human study is also done to evaluate a sampled subset of the course recommendations to compare the alignment between the human and LLM evaluations.

5.3.1 Query Templates. To address the major themes of student concerns in course selection identified in RQ1, we design structured prompt templates. Specifically, we focus on the following categories:

(1) **Course Suitability and Relevance:** Example - "I want to work in _____. What advanced courses should I take?"
(2) **Program and Graduation Requirements:** Example - "What courses do I still need to take to complete my _____ program?"
(3) **Workload and Performance Concerns:** Example - "What are the easiest _____ courses to boost my GPA?"

These prompt templates are further customized by incorporating keywords tailored to specific programs of study. We consider four comprehensive profiles:

(1) **Social Sciences** (e.g., Economics, Sociology, Political Science)
(2) **Arts and Humanities** (e.g., History, Cinema Studies, Philosophy)
(3) **STEM** (e.g., Computer Science, Engineering, Biology)
(4) **Professional and Applied Studies** (e.g., Business Administration, Education, Nursing)

For each profile, we select 12 keywords, reflecting common terms and concerns in that domain (e.g. Political Science for Social Sciences). This approach ensures diversity and relevance in evaluation.

5.3.2 Evaluation Criteria. To evaluate the LLM's ability to recommend courses effectively, we assessed its performance using five metrics (scored from 1 to 5) designed to capture the system's alignment with user goals and its ability to address diverse needs. These metrics are used by the LLM evaluator and the human evaluators to score the LLM course recommendations:

- **Relevance**: How well do the recommendations align with the user's query, interests, and goals.
- **Personalization**: To what degree are the recommendations tailored to the user's preferences and background.
- **Flexibility**: Whether the system adapts to user-specific constraints such as scheduling or workload preferences.
- **Transparency**: Whether the system provides sufficient justification or rationale for its recommendations.
- **Exploration Encouragement**: Whether the system encourages users to explore new or less obvious course options.

For each category and each profile, we generate N=10 prompts by sampling from the keywords. This results in 120 course recommendations, which are evaluated by the LLM evaluator based on the Evaluation Criteria. Additionally, a subset of n=5 course recommendations is randomly sampled and independently evaluated by two experimenters to ensure validity and consistency.

5.4 Results

Tables 3, 4, and 5 summarize the evaluation outcomes of the LLM-based course recommendation system. Across all academic profiles—Social Sciences, Arts and Humanities, STEM, and Professional

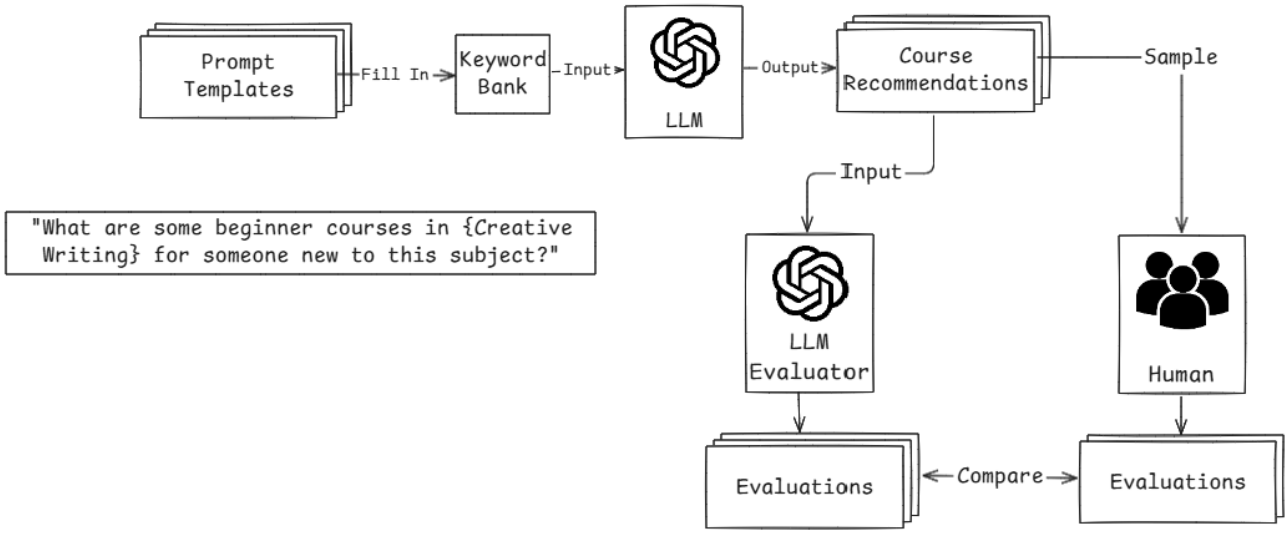

Figure 6: Overview of Evaluation: (1) A set of prompt templates is first customized to diverse student profiles to evaluate different use cases (example provided in block). (2) Templates are used to prompt the LLM course recommender to produce recommendations, which are (3) evaluated by an LLM evaluator according to the five metrics. (4) A subset of the recommendations is sampled and evaluated by human (experimenters) to validate LLM evaluations.

Table 3: LLM-Based Evaluation Results (LLM Only)

Profile	Relevance	Personalization	Flexibility	Transparency	Exploration	Average
Social Sciences	4.7	4.5	4.2	4.6	4.1	4.42
Arts and Humanities	4.5	4.3	4.4	4.5	4.0	4.34
STEM	4.6	4.4	4.3	4.7	4.2	4.44
Professional and Applied Studies	4.4	4.5	4.1	4.3	4.0	4.26
Overall Average	**4.55**	**4.43**	**4.25**	**4.53**	**4.08**	**4.37**

Table 4: Evaluation by Categories of Student Needs (LLM Only)

Category	LLM Score
Course Suitability and Relevance	4.7
Program and Graduation Requirements	4.5
Workload and Performance Concerns	4.4
Overall Average	**4.4**

and Applied Studies—ratings for Relevance, Personalization, Flexibility, Transparency, and Exploration Encouragement were consistently high, averaging between 4.26 and 4.44 on a 5-point scale (Table 3). The system performed strongest in Relevance (overall average 4.55) and Transparency (overall average 4.53), indicating that it effectively identified courses aligned with users' stated goals and provided clear rationales for its suggestions. Although still favorable, Exploration Encouragement showed relatively lower scores (overall average 4.08), suggesting slightly less emphasis on prompting users to consider a broader range of options beyond their initial interests.

When broken down by categories of student needs (Table 4), the LLM's recommendations excelled in Course Suitability and Relevance (4.7), followed by strong performance in meeting Program and Graduation Requirements (4.5), and addressing Workload and Performance Concerns (4.4). These category-level results indicate that the LLM recommendations are well-aligned with users' academic and practical constraints, such as prerequisites and workload management.

Validation by a small sample of human experimenters closely mirrored these LLM-generated assessments (Table 5). Differences between LLM scores and experimenter ratings were minimal (within ±0.05 for most metrics), resulting in nearly identical overall averages (4.37 vs. 4.38). This close alignment suggests that the LLM's scoring methodology and final evaluations are consistent with human judgment.

6 Discussion

The results demonstrate that the LLM-based recommendation system provides high-quality, relevant course suggestions well-aligned with users' academic interests and requirements. Its strong scores in Relevance and Transparency suggest that users receive clear and

Table 5: Validation of LLM Results by Experimenters (Small Human Sample)

Criteria	LLM Score (Self-Generated)	Experimenter Rating	Difference
Relevance	4.55	4.6	-0.05
Personalization	4.43	4.4	+0.03
Flexibility	4.25	4.3	-0.05
Transparency	4.53	4.5	+0.03
Exploration Encouragement	4.08	4.1	-0.02
Overall Average	**4.37**	**4.38**	**-0.01**

accurate justifications for the recommended courses, potentially increasing trust and user satisfaction. The high Relevance rating also indicates that the system effectively interprets user goals and tailors suggestions accordingly, while its robust Transparency scores highlight the system's capability to explain why certain courses are well-suited to the user's stated objectives.

Notably, the system's performance on encouraging exploration, though still positive, was somewhat lower compared to other metrics. This finding suggests that while the system excels at matching users with courses closely fitting their immediate interests and needs, it could benefit from strategies that gently prompt students to consider a more diverse array of academic options. For instance, incorporating logic that introduces complementary subjects, interdisciplinary seminars, or emerging fields in political science may enrich the user's academic journey and foster intellectual curiosity.

The category-level evaluations further emphasize the system's strengths in meeting practical academic concerns, such as ensuring that recommended courses align with program requirements and balancing workload considerations. However, there may still be opportunities to improve personalization by incorporating even more user-specific factors, such as previously completed coursework, learning styles, or long-term career aspirations. Strengthening this personalization could involve refining underlying language models, integrating more detailed institutional data, or including user feedback loops.

The validation exercise with human experimenters confirms the credibility of the automated assessments. The close agreement between LLM-based and human scores supports the reliability of the evaluation framework and suggests that these automated methods could be employed at scale to assist students in their course selection processes.

7 Limitations and Future Works

The LLM-based recommendation system is designed and tested using undergraduate courses offered at the University of Toronto, which may not directly generalize to other course datasets with different academic structures, course offerings, or program requirements. The current system may not support more complex constraints and student interests, such as scheduling issues or financial considerations. Although the model excels at identifying relevant and transparent recommendations, it is less effective in encouraging students to explore broader or less obvious course options.

Future work should focus on incorporating richer user data, such as previous coursework, declared specialties, and availability of real-time courses, to increase personalization and practical applicability. In addition, refining the exploration features of the system, for example, by offering thematic pathways or highlighting interdisciplinary connections, could help broaden students' academic horizons. Finally, validating the system with a diverse pool of actual users and incorporating direct feedback will be crucial to ongoing improvement and adoption in real educational contexts.

8 Conclusions

This study successfully demonstrates the potential of LLM-powered recommendation systems to address the multifaceted challenges students face in course selection. Through an analysis of 247 student discussions, the research identified five key challenges: (1) Course Suitability and Relevance, where students struggle to find courses aligning with their academic and career goals; (2) Program and Graduation Requirements, with students seeking clarity on prerequisites and mandatory course requirements; (3) Scheduling and Accessibility, including conflicts and limited remote learning options; (4) Workload and Performance Concerns, reflecting the desire for manageable workloads or "easy" courses to optimize GPA; and (5) Instructor and Teaching Quality, highlighting the importance of pedagogical effectiveness.

The proposed LLM course recommendation framework effectively addresses these challenges by interpreting natural language queries, providing tailored recommendations, and ensuring transparency in decision-making. Evaluation results indicate strong performance in relevance and transparency, confirming the system's ability to align with students' goals and justify recommendations.

References

[1] P. Arcidiacono, E. M. Aucejo, and K. I. Spenner. 2012. What happens after enrollment? An analysis of the time path of racial differences in GPA and major choice. *IZA Journal of Labor Economics* 1, 5 (2012), 1–24. https://doi.org/10.1186/2193-8997-1-5

[2] P. Attewell and D. Monaghan. 2016. How Many Credits Should an Undergraduate Take? *Research in Higher Education* 57, 6 (2016), 682–713. https://doi.org/10.1007/s11162-015-9401-z

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Gaurav Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[5] A. M. Durik, C. M. Lovejoy, and S. J. Johnson. 2009. A longitudinal study of achievement goals for college in general: Predicting cumulative GPA and diversity in course selection. *Contemporary Educational Psychology* 34, 2 (2009), 113–119. https://doi.org/10.1016/j.cedpsych.2008.11.002

[6] Ahmed Elbadrawy and George Karypis. 2016. Domain-aware grade prediction and top-n course recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 183–190.

[7] T. Gong, J. Li, J. Y. Yeung, and X. Zhang. 2024. The association between course selection and academic performance: exploring psychological interpretations.

Studies in Higher Education 49, 12 (2024), 2506–2519. https://doi.org/10.1080/03075079.2024.2310151

[8] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2024. Large Language Models for Generative Recommendation: A Survey and Visionary Discussions. arXiv:2309.01157 [cs.IR] https://arxiv.org/abs/2309.01157

[9] L. McKinney, A. B. Burridge, M. M. Lee, G. V. Bourdeau, and M. Miller-Waters. 2022. Incentivizing Full-Time Enrollment at Community Colleges: What Influences Students' Decision to Take More Courses? *Community College Review* 50, 2 (2022), 144–170. https://doi.org/10.1177/00915521211061416

[10] Eni Mustafaraj and Panagiotis T Metaxas. 2010. From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search. In *Web Science Conference (WebSci10)*.

[11] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. A Survey on Large Language Models for Recommendation. arXiv:2305.19860 [cs.IR] https://arxiv.org/abs/2305.19860

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009