machine learning & knowledge extraction

MDPI

Article

Leveraging LLMs for Automated Extraction and Structuring of Educational Concepts and Relationships

Tianyuan Yang [1], Baofeng Ren [1], Chenghao Gu [1], Tianjia He [1], Boxuan Ma [2] and Shin'ichi Konomi [2,*]

[1] Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan; yang.tianyuan.791@s.kyushu-u.ac.jp (T.Y.); ren.baofeng.817@s.kyushu-u.ac.jp (B.R.); gu.chenghao.564@s.kyushu-u.ac.jp (C.G.); tianjiahe189@gmail.com (T.H.)
[2] Faculty of Arts and Science, Kyushu University, Fukuoka 819-0395, Japan; boxuan@artsci.kyushu-u.ac.jp
* Correspondence: konomi@artsci.kyushu-u.ac.jp; Tel.: +81-92-802-5875

Abstract

Students must navigate large catalogs of courses and make appropriate enrollment decisions in many online learning environments. In this context, identifying key concepts and their relationships is essential for understanding course content and informing course recommendations. However, identifying and extracting concepts can be an extremely labor-intensive and time-consuming task when it has to be done manually. Traditional NLP-based methods to extract relevant concepts from courses heavily rely on resource-intensive preparation of detailed course materials, thereby failing to minimize labor. As recent advances in large language models (LLMs) offer a promising alternative for automating concept identification and relationship inference, we thoroughly investigate the potential of LLMs in automatically generating course concepts and their relations. Specifically, we systematically evaluate three LLM variants (GPT-3.5, GPT-4o-mini, and GPT-4o) across three distinct educational tasks, which are concept generation, concept extraction, and relation identification, using six systematically designed prompt configurations that range from minimal context (course title only) to rich context (course description, seed concepts, and subtitles). We systematically assess model performance through extensive automated experiments using standard metrics (Precision, Recall, F1, and Accuracy) and human evaluation by four domain experts, providing a comprehensive analysis of how prompt design and model choice influence the quality and reliability of the generated concepts and their interrelations. Our results show that GPT-3.5 achieves the highest scores on quantitative metrics, whereas GPT-4o and GPT-4o-mini often generate concepts that are more educationally meaningful despite lexical divergence from the ground truth. Nevertheless, LLM outputs still require expert revision, and performance is sensitive to prompt complexity. Overall, our experiments demonstrate the viability of LLMs as a tool for supporting educational content selection and delivery.

Keywords: AI in education; concept generation; large language models; relation identification; prompt engineering

Academic Editor: Karin Verspoor

Received: 1 August 2025
Revised: 11 September 2025
Accepted: 15 September 2025
Published: 19 September 2025

Citation: Yang, T.; Ren, B.; Gu, C.; He, T.; Ma, B.; Konomi, S. Leveraging LLMs for Automated Extraction and Structuring of Educational Concepts and Relationships. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 103. https://doi.org/10.3390/make7030103

1. Introduction

In MOOC environments, learners often have the autonomy to select courses based on their interests and educational objectives. However, the vast array of available courses can make it challenging for students to identify the most suitable courses to satisfy their diverse needs [1]. To support informed decision-making, it is essential to provide students with the

core concepts of a course and the relationships between the concepts. Such information offers valuable insights into the course content and the prerequisites necessary for effective learning [2]. While educational institutions and MOOC platforms often provide a sharing environment for course materials, syllabi, and keywords, faculty members or academic staff have to create these resources manually. This process is both time-consuming and resource-intensive, posing a significant challenge to scalability [3].

To address this, researchers have focused on the automatic extraction of course concepts [3,4] and the relations between these concepts [5,6] using course information. For example, Lu et al. [4] proposed DS-MOCE, which leverages pre-trained language models and discipline-specific embeddings to extract course concepts from MOOCs with minimal manual annotation. Aytekin et al. [5] presented a machine learning-assisted framework that integrates semantic analysis and expert validation to generate concepts with prerequisite relations. While these approaches effectively identify concepts and their relationships, they exhibit several notable limitations. One major challenge is their heavy reliance on detailed course content [4]. Existing studies typically extract course concepts from textual materials and predict relationships based on metrics such as the location and frequency of concepts within the text, making it difficult to generate high-quality course concepts when limited information is available [3]. Therefore, these methods focus on explicit textual features rather than conceptual inference, and they struggle to generate concepts that may not appear in the text or that occur infrequently, even though these concepts are crucial for understanding the course [3]. For example, consider the following course description: "This course introduces supervised and unsupervised learning, covering linear regression, logistic regression, decision trees, and K-means clustering". Previous methods would identify concepts like supervised learning, unsupervised learning, linear regression, and K-means clustering. However, they would likely fail to recognize important but unstated concepts such as Bayesian classification, model evaluation, overfitting, and data preprocessing, which are essential to understanding machine learning. Similarly, the identification of inter-conceptual relationships is highly constrained by the availability of conceptual information [7]. Previous methods often rely on explicit co-occurrences or external knowledge bases, making them tend to identify only surface-level associations rather than capturing deeper semantic or causal relationships between concepts.

The integration of Large Language Models (LLMs), such as GPT, into educational practices has garnered increasing attention as educational paradigms evolve and technology-driven approaches gain prominence. With their ability to understand context, generalize beyond literal content, and infer implicit relationships, LLMs have the potential to overcome key limitations of traditional NLP methods. Recent studies have explored the use of LLMs for generating course-related content, including knowledge concepts and syllabi [8–11]. However, despite these promising capabilities, LLM-generated outputs can sometimes include factual inaccuracies or logical inconsistencies. While such issues may be relatively easy to identify in general tasks like essay writing or programming, they become significantly harder to detect in more specialized educational tasks, such as the generation of knowledge concepts or prerequisite structures [12,13]. These subtleties pose challenges for quality assurance and highlight the need for rigorous and systematic evaluation of LLMs' performance in educational contexts, especially in tasks that involve curriculum-level concept modeling and relation identification. This challenge underscores the need for systematic evaluation of AI-generated outputs to ensure their reliability and educational value. However, none of these studies have systematically evaluated LLMs' ability to generate and extract curriculum concepts, let alone explored their potential for identifying inter-conceptual relationships. Existing LLM-based studies in education have primarily been exploratory and narrow in scope. For instance, Yang et al. [9] examined the feasibility of us-

ing GPT to expand course concepts, but without systematic benchmarking or cross-model comparison. Similarly, Ehara [10] only compared GPT-generated concepts with manual annotations, while other efforts have focused on qualitative coding [14] or tutoring support rather than curriculum-level concept modeling. These works demonstrate the potential of LLMs but remain fragmented, typically addressing single tasks and lacking rigorous evaluations that combine automated metrics, statistical analysis, and expert validation. In contrast, our study provides the first systematic and comparative evaluation of multiple LLM variants across concept generation, concept extraction, and relation identification, under six carefully designed prompt configurations. This design enables us to highlight when and how LLMs can reliably support educational applications.

Motivated by this gap, we conduct a systematic evaluation of LLMs' ability to generate course concepts and identify their inter-conceptual relationships. This paper explores the feasibility of applying LLMs in the educational domain, with a particular focus on their ability to generate relevant concepts and relations based on course information. To comprehensively assess the capability of LLMs in educational concept identification, we examine both concept generation and concept extraction. Concept generation represents an open-ended task that explores the breadth of concepts LLMs can propose, while concept extraction provides a constrained setting that allows for corrective filtering and precision. Studying both tasks together enables us to capture complementary perspectives on the strengths and limitations of LLMs. To this end, we conduct a comprehensive evaluation of LLM-based models across three core tasks: concept generation, concept extraction, and relation identification. Our experiments systematically vary the input granularity through six tailored prompts and evaluate three LLM variants (GPT-3.5, GPT-4o-mini, GPT-4o). We assess performance using both automatic metrics and human expert evaluations, comparing LLM-generated outputs not only with traditional NLP baselines, but also against human-annotated ground truth from a real-world dataset. Our study aims to address the following key research questions:

RQ1: How do different input prompts affect LLMs' performance in course concept generation and extraction across varying levels of contextual detail?
RQ2: To what extent can LLM-generated course concepts align with human-annotated ground truth and outperform traditional NLP baselines in terms of quality and coverage?
RQ3: Can LLMs accurately infer prerequisite relationships between course concepts, and how does their performance vary under different information conditions?
RQ4: Can we use LLM-generated concepts and relations for supporting practical educational scenarios?

This study makes the following contributions:

- We systematically evaluate LLMs across three core educational tasks: concept generation, concept extraction, and relation identification.
- We design six levels of prompt configurations that vary in informational granularity, enabling fine-grained analysis of how input context affects LLMs' performance across different task constraints.
- We show the effectiveness and reliability of LLM-generated outputs through comprehensive quantitative and qualitative evaluations, including comparisons with traditional NLP baselines and human expert assessments.
- We demonstrate the practical value of LLM-generated concepts and relations in educational scenarios such as course metadata enrichment, knowledge graph construction, and prerequisite-aware course recommendation.

By addressing these questions, this study contributes to the growing body of knowledge on the application of AI technologies in education. Our results demonstrate that

LLMs can generate high-quality knowledge concepts and accurate inter-conceptual relations across both prompt styles. Furthermore, detailed course information enhances LLMs' ability to produce more standardized and higher-quality knowledge structures. This approach not only saves significant time and effort compared to manual construction but also provides students with a more efficient way to understand and select courses. Moreover, it offers valuable insights for the development of AI-driven educational tools, such as course recommendation systems, paving the way for more effective and scalable solutions in the education sector.

2. Related Work

2.1. Concept Extraction and Relation Identification

Educators have long recognized the critical role of concepts and their prerequisite relationships in learning resources, as these are essential for helping students understand the curriculum and select suitable courses. Significant progress has been made in automating the identification of key concepts and relations within educational materials.

Early efforts to automate course concept extraction employed a range of semi-supervised [15], embedding-based [3], and graph-driven techniques [16]. While effective to some extent, these approaches often suffered from scalability limitations, heavy reliance on textual content, or vulnerability to semantic noise. Foster et al. [15] proposed a semi-supervised learning approach for core concept identification using expert-annotated features, but its reliance on labeled data limited scalability. Similarly, Changuel et al. [17] tackled identifying effective learning paths from web document corpora by annotating results and prerequisite concepts. Pan et al. [3] proposed a method to extract and rank fine-grained course concepts in MOOCs using embedding-based representations and a graph-based propagation algorithm, addressing challenges such as low-frequency concepts in video captions. Manrique et al. [16] applied knowledge graphs to rank concepts, yet their approach was constrained by entity-linking quality and knowledge completeness. Yu et al. [18] expanded course concepts using external knowledge bases and interactive feedback, but their method suffered from semantic drift and noise. Although these methods effectively address concept extraction, they rely heavily on textual data and often incur high computational costs due to complex models.

Given the importance of prerequisite relationships between concepts, numerous studies have focused on this area, although extracting such relationships from textual data remains challenging. Many researchers have relied on unsupervised or supervised techniques to detect prerequisite relationships, particularly between Wikipedia articles [16,19]. Liang et al. [20] proposed Reference Distance (RefD), a link-based metric that utilizes Wikipedia hyperlinks to assess prerequisite relations. Pan et al. [6] introduced embedding-based methods to identify relationships in MOOCs, leveraging textual data for relational inference. Yet, this method still heavily relies on explicit textual cues, failing to effectively infer implicit concept relations that are common in many curricula. Manrique et al. [21] explored the use of general-purpose knowledge graphs, such as DBpedia and Wikidata, to model concept dependencies. Li et al. [22] presented LectureBank, a dataset of 1352 lecture files for NLP and related domains, and explored prerequisite chain learning using graph-based neural networks and traditional classifiers. Zhang et al. [23] developed a variational graph autoencoder designed to estimate precedence relations within knowledge graphs. More recently, Aytekin et al. [5] proposed ACE, a machine learning-assisted approach that integrates expert feedback to construct Educational Knowledge Graphs, significantly reducing the need for manual labeling.

While these methods have shown effectiveness in concept extraction and relationship identification, they exhibit several limitations. Notably, they heavily depend on textual

information from course materials and are unable to generate insights beyond the provided text. Furthermore, improved performance often comes at the cost of increased computational overhead and more complex model architectures. Recent work has integrated large language models (LLMs) and knowledge graphs to enhance concept identification [9,10,24]. However, few studies have systematically evaluated LLMs' capabilities in concept generation and relation identification tasks.

2.2. Large Language Models in Education

LLMs pre-trained on extensive textual data have become a cornerstone of modern NLP research. Recent advancements in NLP have led to the development of high-performing LLMs, such as GPT-3, GPT-4, and Claude, which excel in tasks like machine translation, text summarization, and question-answering [25]. Furthermore, studies have demonstrated that LLMs can achieve remarkable results in downstream tasks with minimal or no demonstrations in the prompt [26–28]. The emergence of LLMs, such as GPT, introduces new educational opportunities, including the automatic generation of educational content, personalized learning experiences, and the enhancement of educational tools [29].

A growing body of research has explored diverse educational applications of LLMs [11,30–32], such as course recommendation, content creation, and addressing data sparsity [10,29]. For instance, Yang et al. [9] utilized GPT to expand course concepts, evaluating the feasibility of using GPT-generated concepts as a direct educational resource. Similarly, Ehara [10] examined the effectiveness of GPT-generated concepts in enhancing interpretability and found that while these concepts aligned well with standard course content, they required further refinement to address inconsistencies. Barany et al. [14] investigated GPT's potential for qualitative codebook development, comparing manual, automated, and hybrid approaches to evaluate their impact on code quality, reliability, and efficiency in educational research. Castleman and Turkcan [33] investigated the integration of knowledge bases into LLM-based tutoring systems, finding that enhanced access to knowledge bases improved these systems' comprehension and communication capabilities, though they still fell short of human expertise. Lin et al. [34] explored GPT-generated feedback for tutor training, aiming to improve educational tool quality. Beyond direct applications, LLMs can also be incorporated into various educational tools, such as course recommendation systems [31,32,35].

While numerous studies have explored the application of LLMs' generative capabilities in education, their focus has largely been limited to specific and small-scale tasks. For example, Yang et al. [9] used GPT to expand course concepts without systematic benchmarking, Ehara [10] only measured similarity between GPT-generated and manual concepts, and Barany et al. [14] investigated GPT for qualitative coding or tutoring support rather than curriculum-level concept modeling. These efforts remain fragmented, lack cross-model comparisons, and do not incorporate rigorous statistical or expert validation. In contrast, our work provides the first systematic and comparative evaluation of multiple LLMs across three fundamental tasks, which are concept generation, concept extraction, and relation identification, under six carefully designed prompt configurations, with both automated metrics and expert assessments. These efforts remain fragmented and primarily exploratory. To move beyond such piecemeal investigations, our work establishes a systematic benchmark that evaluates multiple LLMs across tasks and prompts with rigorous automated and expert-based assessments.

3. Methodology

We conducted our experiments on the MOOCCube dataset [36], which contains 683 courses, 25,161 unique concepts, and 1027 prerequisite relations. Each course includes

textual descriptions, subtitles, and manually annotated concepts. For evaluation, we randomly sampled 100 courses across multiple domains; this same subset was used for all models and baselines to ensure comparability. The overall workflow for utilizing LLMs for concept generation and relation identification is illustrated in Figure 1. Specifically, we developed two concept-level tasks, concept generation and concept extraction, as well as a relation-level task, relation identification. These tasks aim to evaluate the performance of LLMs in both concept-level and relation-level tasks. To facilitate these tasks, we designed specific prompts for LLMs to generate the required concepts and relations.

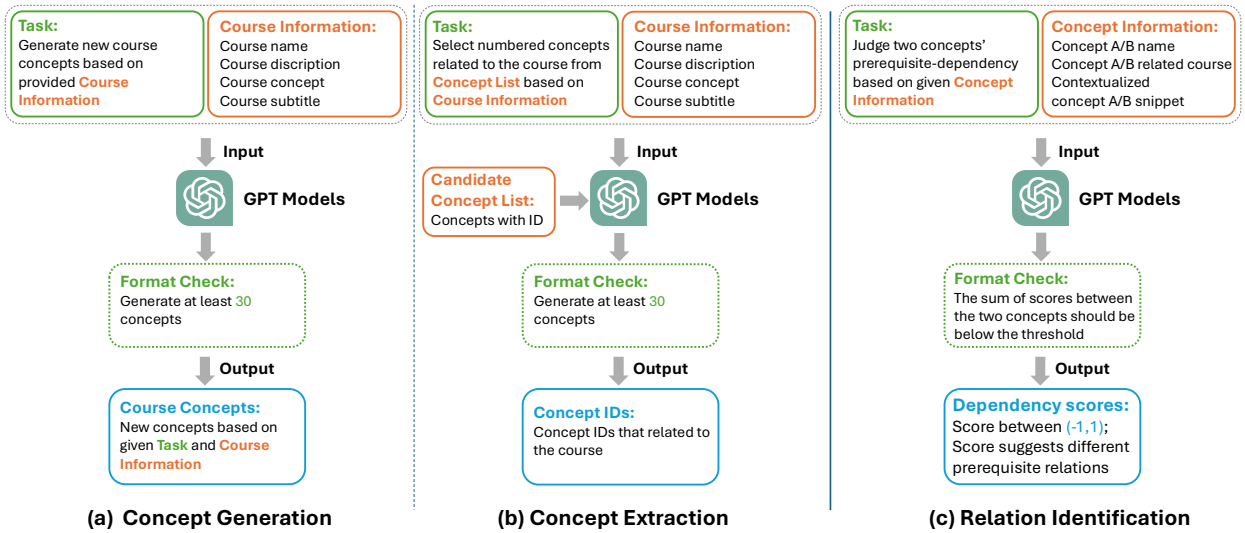


Figure 1. Workflow of utilizing LLMs to perform two concept-level tasks and one relation-level task. (**a**,**b**) are concept-level tasks, while (**c**) is relation-level task [37].

3.1. Concept-Level Task Design

3.1.1. Concept Generation

The concept generation task allows LLMs to produce outputs based on the input prompt without strict constraints or predefined response ranges. This task leverages the model's semantic understanding and generative capabilities to generate diverse and potentially innovative responses. Previous studies on concept generation tasks have primarily focused on open-ended generation tasks [24,38]. As shown in Figure 1a, we utilize LLMs to generate relevant concepts for each course. We leverage the target course and its related information and incorporate this information to construct a prompt for LLMs. The prompt for the concept generation task consists of three components: task description, format indicator, and information injection. As illustrated in Figure 2, the task description specifies the objective, such as concept generation, and outlines the required information. The format indicator defines the desired output structure and the number of concepts to be generated. To ensure LLMs meet the specified requirements where LLMs may overlook parts of the prompt requirements, a retry mechanism is implemented to verify that the output aligns with the prompt's criteria. Information injection provides LLMs with relevant course details, including course name, description, related concepts, video subtitles, and examples.

To systematically assess the impact of varying information granularity on LLMs' performance in generating course concepts, we designed six prompts by ablating the information injection: *Zero-Shot* (P1), *One-Shot* (P2), *Concept* (P3), *Desc* (P4), *Concept+Desc* (P5), and *Subtitle* (P6) to provide LLMs with different levels of contextual information. As shown in Table 1, *Zero-Shot* (P1) and *One-Shot* (P2) offer minimal course information,

with *Zero-Shot* (P1) providing only the course name and *One-Shot* (P2) including both the course name and an example. *Concept* (P3), *Desc* (P4), and *Concept+Desc* (P5) supply more general course-related details, with *Concept* (P3) providing the course name, an example, and related concepts; *Desc* (P4) offering the course name, an example, and a course description; and *Concept+Desc* (P5) combining the course name, an example, related concepts, and the course description. *Subtitle* (P6) delivers comprehensive course information by extending *Concept+Desc* (P5) to incorporate subtitle information from the course video. After designing these prompts to generate course knowledge concepts using LLMs, we successfully generated relevant concepts for 100 courses, with each course having a rich and comprehensive set of generated concepts. By structuring the prompts in this manner, we aim to evaluate how varying levels of information granularity influence LLMs' ability to generate relevant course concepts.

Figure 2. Examples of concept generation and extraction task prompts [37].

Table 1. Information injection components across prompts (P1–P6) in concept generation and extraction tasks.

Information	P1 Zero-Shot	P2 One-Shot	P3 Concept	P4 Description	P5 Concept+Desc	P6 Subtitle
Course Name	✓	✓	✓	✓	✓	✓
Example		✓	✓	✓	✓	✓
Related Concepts			✓		✓	✓
Course Description				✓	✓	✓
Video Subtitles						✓

3.1.2. Concept Extraction

The concept extraction task, distinct from concept generation, requires LLMs to select the most appropriate concepts from a set of predefined options provided in the input

prompt. The model must comprehend the options and make accurate, context-based selections. This task emphasizes LLMs' inference and context-based judgment abilities. To comprehensively assess LLMs' selection performance, we developed a novel method and corresponding prompts. The overall workflow of the proposed method for the concept extraction task is shown in Figure 1b. The difference in the concept extraction task is that we provide a pre-defined and pre-numbered candidate concept list, as specified in the prompt, and require LLMs to select the concepts from this list that are most suitable for the course. Notably, when designing the prompt, we pre-numbered the concept list and instructed LLMs to output the numbers corresponding to the concepts relevant to the course, rather than the concepts themselves to prevent *hallucination*. This approach was adopted because, during testing, LLMs were asked to output concepts from the candidate list directly occasionally *hallucinated* and generated concepts not present in the provided list. By pre-numbering the concepts and having LLMs output the corresponding numbers, we ensured that the generated concepts strictly belonged to the original candidate list. The prompt design can be found in the lower section of Figure 2. The prompt for the concept extraction task consists of four components: a precise *task description* clearly instructing the selection from a given candidate list, strictly prohibiting concept generation; a detailed *format indicator* specifying output only as numerical identifiers corresponding to candidate concepts, effectively preventing LLMs' hallucinations; an explicit *information injection*, similar in form to that in concept generation, but differing substantially in its role—it serves as contextual reference to guide accurate selection rather than inspiration for new concept creation; and the carefully constructed *candidate concept list*, pre-numbered to align with the numerical output format required by LLMs.

In parallel with the generation task, we designed six analogous prompts for extraction, systematically varying informational granularity: *Zero-Shot* (P1), *One-Shot* (P2), *Concept* (P3), *Desc* (P4), *Concept+Desc* (P5), and *Subtitle* (P6), to provide LLMs with different levels of contextual information. Each incorporating different combinations of course information are detailed in Table 1. After designing prompts for extracting course knowledge concepts using LLMs, we successfully extracted concepts for 100 courses. These complementary concept-level tasks offer a comprehensive methodological framework to critically evaluate LLMs' performance, particularly their generative creativity and constrained reasoning skills, in educational content scenarios.

3.2. Relation-Level Task Design

Relation identification involves identifying and extracting semantic relations between concepts from curriculum content, descriptive texts, or knowledge structures. The goal is to elucidate associations among course concepts such as *prerequisite relations*, *similarity relations*, and *containment relations* to facilitate the construction of knowledge graphs, improve course recommendations, and optimize student learning paths in educational contexts. Aligned with previous studies, we focus on identifying prerequisite relations [5,22]. The overall workflow, as illustrated in Figure 1c, we fed a pair of concepts (*Concept A*, *Concept B*) into LLMs with specifically designed prompts containing varying levels of information about the concepts. LLMs output a numerical score in the range of -1 to 1, representing the likelihood of a prerequisite relationship between Concept A and Concept B in the pair. We adopt the $[-1, 1]$ range following the design of Reference Distance (RefD) [20], which encodes prerequisite directionality on a symmetric interval. In this formulation, a positive score indicates that Concept A is likely a prerequisite of Concept B, a negative score indicates the reverse direction, and values near zero denote the absence of a prerequisite relation. Concretely, the LLM is prompted to provide two directional plausibility scores, $r_{A\to B}$ and $r_{B\to A}$, each in $[0, 1]$ under a mutual-exclusion constraint. We then compute a signed score

$s(A, B) = r_{A \to B} - r_{B \to A} \in [-1, 1]$, which compactly encodes both directionality and confidence. To obtain discrete labels for evaluation, we apply a thresholding step with margin θ: if $s(A, B) > \theta$, we assign $+1$ (A is a prerequisite of B); if $s(A, B) < -\theta$, we assign -1 (B is a prerequisite of A); otherwise, we assign 0 (no relation). This threshold-based discretization ensures that uncertain cases are conservatively mapped to "no relation". Unlike cosine similarity, which measures vector closeness, our signed scale explicitly enforces asymmetry and supports graded confidence while producing categorical judgments for downstream analysis. The prompt design is illustrated in Figure 3, which comprises four key components: *task description*, *prior knowledge*, *format indicator*, and *information injection*. The task description instructs LLMs to output likelihood scores for two scenarios: Concept A as a prerequisite for Concept B, and vice versa. We designed the prompt based on [20], which defines prerequisite relationships by stating that if Concept A is a prerequisite for Concept B, the reverse cannot be true. We incorporated this as prior knowledge, ensuring that the likelihood scores of reversed relationships ideally approach zero. To enforce this asymmetry, the format indicator includes a retry mechanism to ensure compliance and specifies that output scores range from -1 to 1, where 1 indicates Concept A is a prerequisite for Concept B, and -1 suggests the opposite. This design effectively mitigates the LLMs' hallucination issue, ensuring more reliable outputs. The information injection component provides LLMs with relevant details about the concept pairs, including their names, contextualized concept snippets as explanations, course names, course descriptions containing the concepts, and relevant examples.

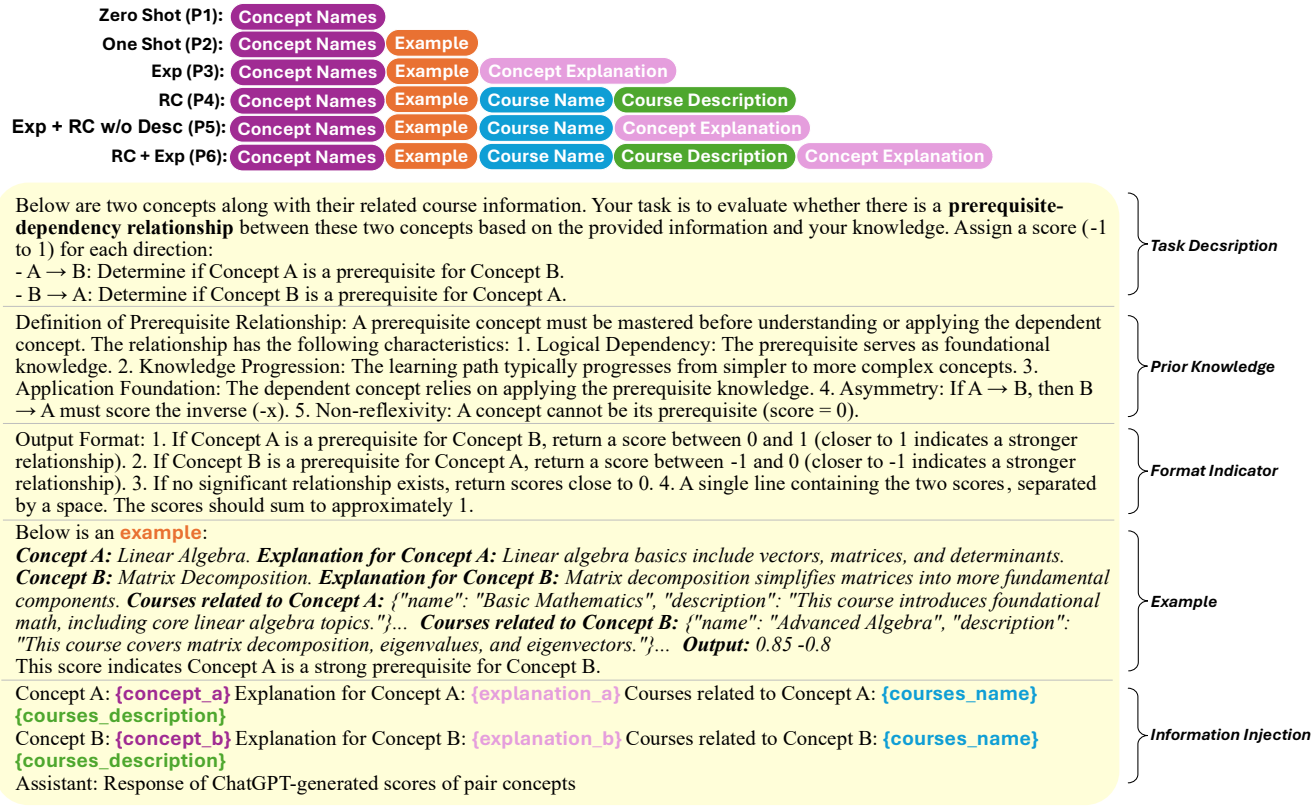


Figure 3. Examples of relation identification task prompts [37].

Specially, we designed six prompts with varying levels of information by ablating the information injection component: *Zero-Shot* (P1), *One-Shot* (P2), *Explanation* (P3), *Related Course* (P4), *Explanation+Related Course w/o Desc* (P5), and *Explanation+Related Course* (P6). The *Zero-Shot* (P1) and *One-Shot* (P2) prompts follow the same structure as in previous tasks, with the *Zero-Shot* prompt providing only the concept pair names, while *One-Shot*

adding an example to illustrate the concept relationship. *Explanation* (P3) builds on P2 by incorporating contextualized snippets that explain each concept separately. *Related Course* (P4) includes course names and descriptions associated with the given concepts based on P2. *Explanation+Related Course* (P6) integrates both contextualized concept snippets and course information, whereas *Explanation+Related Course w/o Desc* (P5) is similar to P6 but excludes course descriptions. Each prompt incorporating different concept information is detailed in Table 2. The rationale behind our prompt design is to explore different levels of information granularity. Specifically, P1 and P2 serve as baselines without additional information. P3 introduces contextual explanations of individual concepts, reflecting a concept-level augmentation, while P4 introduces related course information, representing a course-level augmentation. Finally, P5 and P6 combine both concept-level and course-level information to examine their joint impact on prerequisite prediction. Utilizing these carefully designed prompts, we successfully evaluated 100 pairs of concepts to determine their prerequisite relationships and generated corresponding scores. Our prompt design effectively mitigates the LLMs' hallucination issue, ensuring more reliable outputs.

Table 2. Information injection components across prompts (P1–P6) in the relation identification task.

Information	P1	P2	P3	P4	P5	P6
Concept Names	✓	✓	✓	✓	✓	✓
Example		✓	✓	✓	✓	✓
Concept Explanation			✓		✓	✓
Course Name				✓	✓	✓
Course Description				✓		✓

4. Experimental Setup

For LLM-based experiments, we used the OpenAI API with a maximum output length of 3000 tokens, a temperature of 0.7, and top_p of 0.9. Since the API does not rely on random seeds, reproducibility is controlled through these fixed parameter settings.

4.1. Dataset

We utilized the MOOCCube dataset [36], a large-scale MOOC dataset collected from the XuetangX platform, to conduct our experiments. Before conducting our experiments, we performed additional preprocessing to ensure data quality and consistency. Specifically, we removed courses containing fewer than 10 associated concepts, as such courses provide insufficient information for reliable evaluation. We also excluded certain courses whose content, such as final year projects or graduation theses from specific universities, could cause semantic ambiguity or misinterpretation. After preprocessing the dataset, it included 683 courses and 25,161 distinct course concepts. Each course in the dataset is associated with a course description and related knowledge concepts. Since each course is presented in video format, the dataset also includes subtitle text corresponding to each course video. An illustrative example of course information in the dataset is shown in Table 3. The concepts provided in the dataset were initially generated using a deep learning model [3] trained to identify key terms from the subtitle text of each course. The model extracted candidate concepts based solely on the textual content of the subtitles. Subsequently, human annotators manually reviewed and refined these outputs to ensure quality and relevance. Additionally, the dataset includes 1027 prerequisite relationships between certain concepts. We used these concepts and relationships as the ground truth for evaluating LLM-generated outputs.

Table 3. An example of course information in the MOOCCube dataset.

Course Name	Manual Concept	Course Description
Principles and Development of Database Systems	Minimum Spanning Tree; Database Technology; Shortest Path	Database technology is a core component of various information systems such as business processing systems, e-commerce systems, management information systems, office automation systems, and big data application systems. It is also a crucial technical means for efficiently managing and utilizing data resources in an information society, supporting business processing, data analysis, information services, scientific research, and decision-making management. The educational objectives of this course are to help learners grasp the principles and development techniques of database systems, and cultivate students' engineering abilities in database design, programming, and innovative applications, thereby establishing their competencies in database application system development.

4.2. Baselines

We compared large language models (LLMs) with traditional NLP methods for course concept generation and extraction tasks. For LLMs, we evaluated three LLM variants: GPT-3.5, GPT-4o-mini, and GPT-4o, chosen for their affordability and widespread adoption. As traditional NLP baselines, we selected three categories of methods for comparison: word frequency-based methods, deep learning-based methods, and graph-based methods. For each category, we selected representative baselines:

- **PMI** [39]: Pointwise Mutual Information measures the statistical association between word pairs based on their co-occurrence, often used to identify strongly related terms.
- **TF-IDF** [40]: TF-IDF assigns weights to terms based on their frequency in a document and rarity across documents, highlighting course-specific keywords.
- **TextRank** [41]: TextRank is a graph-based ranking algorithm that scores terms based on co-occurrence, widely used for unsupervised keyword extraction.
- **Word2Vec** [42]: Word2Vec learns dense word embeddings by modeling context, enabling the identification of semantically similar terms in course texts.
- **BERTScore** [43]: BERTScore evaluates term similarity using contextual embeddings from pre-trained BERT models, capturing deep semantic alignment.
- **TPR** [44]: TPR combines topic decomposition and graph propagation to extract and rank keyphrases, effectively capturing topic-representative concepts.

Among these baselines, PMI, TF-IDF, and TextRank are word frequency-based methods; Word2Vec and BERTScore are deep learning-based methods; and TPR is a graph-based method, following the classification shown in Table 4. For traditional NLP baselines such as PMI, TF-IDF, TextRank, Word2Vec, BERTScore, and TPR, no explicit training–testing split was performed. Each method was applied independently to the subtitles of 100 courses to extract candidate concepts, which were then compared against ground-truth annotations.

Since these approaches operate in an unsupervised manner, the reported results directly reflect their ability to extract relevant terms from course transcripts without additional model training.

Table 4. Performance comparison (%) on MOOCCube dataset [37].

Category	Method	Precision	Recall	F1 Score	Accuracy
Word Frequency	PMI	2.90	0.90	1.26	0.63
	TF-IDF	16.33	2.61	4.35	2.25
	TextRank	14.78	2.07	3.60	1.85
Deep Learning	W2V	12.17	1.56	2.75	1.43
	BERTScore	10.17	1.64	2.70	1.38
Graph-based	TPR	13.50	1.99	3.43	1.76
LLMs	GPT-3.5	**67.48**	**39.32**	**46.38**	**34.03**
	GPT4o-mini	13.97	19.41	15.29	9.08
	GPT4o	17.90	21.37	18.55	12.27

4.3. Evaluation Metrics

We employed four widely adopted evaluation metrics: Precision, Recall, F1 Score, and Accuracy to quantitatively assess the performance of LLM-generated concepts, following prior studies in concept extraction and educational NLP [22,24]. These metrics are commonly used to evaluate the alignment between predicted outputs and ground-truth annotations, providing a comprehensive view of model performance. Higher metric values indicate better alignment and overall effectiveness.

Precision measures the proportion of predicted concepts that are relevant, which can be defined as:

$$\text{Precision} = \frac{\left| C_{\text{pred}} \cap C_{\text{true}} \right|}{\left| C_{\text{pred}} \right|},$$

where C_{pred} denotes the set of predicted concepts, and C_{true} denotes the set of ground-truth concepts. Precision evaluates the correctness of the predicted concepts.

Recall quantifies the proportion of relevant concepts that are successfully predicted, which can be calculated as:

$$\text{Recall} = \frac{\left| C_{\text{pred}} \cap C_{\text{true}} \right|}{\left| C_{\text{true}} \right|},$$

where C_{pred} is the predicted concept set, and C_{true} is the reference set. Recall answers the question of how completely the relevant concepts are retrieved.

F1 Score evaluates the overall agreement between the predicted and ground-truth concept sets, which can be obtained as:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Accuracy measures the overall agreement between the predicted and ground-truth concept sets, and is defined as:

$$\text{Accuracy} = \frac{\left| C_{\text{pred}} \cap C_{\text{true}} \right|}{\left| C_{\text{pred}} \cup C_{\text{true}} \right|},$$

where $C_{pred} \cup C_{true}$ represents the union of predicted and true concept sets. It reflects both precision and recall from a set similarity perspective. In addition to automated evaluation, we also conducted human assessments to complement these quantitative metrics, following previous studies on concept quality evaluation [5,9].

5. Results

5.1. Performance on Concept Generation

5.1.1. Performance Comparison with Baselines

We compare our approach with a set of representative baseline methods from the NLP domain, spanning statistical, graph-based, and embedding-based techniques. All methods were provided with the same input, subtitle transcripts of course videos, corresponding to our prompt configuration *Subtitle* (P6). For consistency, we selected 100 courses from the MOOCCube dataset and generated at least 30 concepts per course using both LLMs and the baselines. Table 4 summarizes the performance results across four evaluation metrics.

Among traditional methods, TF-IDF, TextRank, and TPR demonstrate relatively better performance compared to PMI and embedding-based approaches. However, their overall F1 scores remain below 5%, indicating limited ability to capture the full semantic scope of the course content. These methods are inherently constrained by surface-level lexical patterns and term frequencies. For example, TF-IDF favors frequent but potentially generic terms, while TextRank and TPR rely on co-occurrence graphs that may fail to prioritize pedagogically meaningful concepts. Embedding-based approaches such as Word2Vec and BERTScore slightly improve precision but still fall short in recall and overall alignment with ground-truth concepts.

In contrast, LLMs, particularly GPT-3.5, achieve significantly higher scores across all metrics. GPT-3.5 reaches a precision of 67.48%, a recall of 39.32%, and an F1 score of 46.38%, vastly outperforming all baselines. This performance gap reflects the model's capacity to integrate contextual cues, infer latent concepts, and generalize beyond the literal content of the subtitles. While GPT-4o and GPT-4o-mini yield lower scores than GPT-3.5 on metrics-based evaluation, this discrepancy does not imply inferior concept quality. Upon closer examination, we find that many concepts generated by the GPT-4o variants are pedagogically meaningful, contextually appropriate, and accurately reflect the course content, despite differing in lexical expression or abstraction level from the annotated ground truth. These differences highlight the models' ability to synthesize relevant knowledge beyond surface-level matching. Although GPT-4o is generally regarded as a more advanced model, GPT-3.5 achieved higher quantitative scores on our string-overlap metrics, a seemingly counterintuitive result. Several factors may explain this discrepancy. First, since the MOOCCube ground truth was constructed by model-based extraction followed by human correction, the annotations may retain lexical patterns characteristic of an extraction-style process. Such patterns emphasize explicit keywords or short phrases, which GPT-3.5 tends to reproduce more directly, leading to higher surface-level overlap with the reference set. Second, prompt–model alignment effects likely play a role: differences in training data distribution, tokenization, and stylistic preferences mean that GPT-3.5's lexical choices align more closely with the annotated vocabulary, whereas GPT-4o tends to generate more abstract or pedagogically framed expressions. Third, we also observe a behavioral difference between the two models. GPT-3.5 often directly extracts or replicates keywords from the subtitles, which naturally favors string-matching metrics. In contrast, GPT-4o frequently summarizes and reformulates the content, producing concepts that align more closely with human judgments of pedagogical relevance but diverge lexically from the annotations. For example, GPT-4o often produced semantically adequate but lexically divergent outputs such as *Bayesian inference* instead of the annotated *Bayes theorem*,

which illustrates how string-overlap metrics systematically undervalued its strengths. This combination of factors explains why GPT-3.5 achieves higher metric-based scores, while GPT-4o performs better in human evaluation and produces concepts that are ultimately more meaningful for educational applications.

Although the absolute values of accuracy (34.03%) and precision (67.48%) may appear relatively low, this is expected given the open-ended nature of concept extraction. Unlike conventional classification tasks, the ground truth in MOOCCube contains only a subset of possible valid concepts, causing many semantically appropriate outputs to be penalized by strict string-overlap metrics. As a result, traditional metrics-based evaluation may undervalue semantically relevant but lexically divergent outputs. To address this limitation, we further conduct a human evaluation (Section 5.2), which confirms that LLM-generated concepts are pedagogically meaningful and often outperform ground-truth annotations in relevance and instructional value.

Beyond metric-based superiority, LLMs also exhibit qualitative advantages. Traditional NLP methods are restricted to extracting terms that are explicitly mentioned in the input text. If a relevant concept is rare or entirely absent from the subtitles, these models are unlikely to recover it. LLMs, on the other hand, leverage pre-trained knowledge and language modeling capabilities to infer semantically relevant but implicit concepts. For example, in a machine learning course, traditional methods tend to extract surface terms such as "*gradient descent*" or "*neural networks*," which appear frequently in the subtitles. LLMs, however, can generate higher-level or prerequisite concepts like "*bias-variance trade-off*" or "*Bayesian inference*," even if these are not explicitly stated in the course transcripts. This capacity to synthesize domain-relevant knowledge beyond the observed data highlights LLMs' potential for supporting educational applications where completeness and pedagogical value are critical.

5.1.2. Ablation Study

To contextualize the following human evaluation, it is important to note that GPT-3.5's higher scores on automated metrics largely stem from its tendency to replicate lexical patterns present in the ground truth, which itself may contain residual model-specific phrasing. GPT-4o, by contrast, often summarizes or reformulates the content, producing semantically appropriate and pedagogically meaningful concepts that diverge lexically from the annotations. As a result, GPT-4o is disadvantaged by surface-level string matching but aligns more closely with human judgments of concept quality. To further investigate how varying levels of contextual input and different LLMs affect concept generation performance, we conducted an ablation study involving six prompt configurations (P1–P6) and three LLM variants: GPT-3.5, GPT-4o-mini, and GPT-4o. Each prompt was designed to introduce more course-related information incrementally, ranging from minimal inputs (e.g., course title only) to comprehensive inputs, including course descriptions, existing concepts, and subtitle transcripts. All generated concepts were compared against the ground-truth annotations in the MOOCCube dataset, and the evaluation results are presented in Figure 4.

Our analysis reveals several important findings. First, increasing the richness of input information consistently enhances performance across all LLM variants. Prompts with more detailed content (P5 and P6) lead to higher precision, recall, and F1 scores, suggesting that LLMs effectively leverage contextual cues to identify relevant concepts. Another notable observation is the difference in model behavior under sparse input conditions. GPT-4o and GPT-4o-mini demonstrate relatively stable performance across low-information prompts (P1–P3), indicating robustness in handling minimal input. In contrast, GPT-3.5 exhibits greater variability in these early prompts, suggesting a higher dependence on input com-

pleteness for generating accurate outputs. These patterns may reflect differing sensitivities to contextual cues and the ways in which each model processes incomplete information.

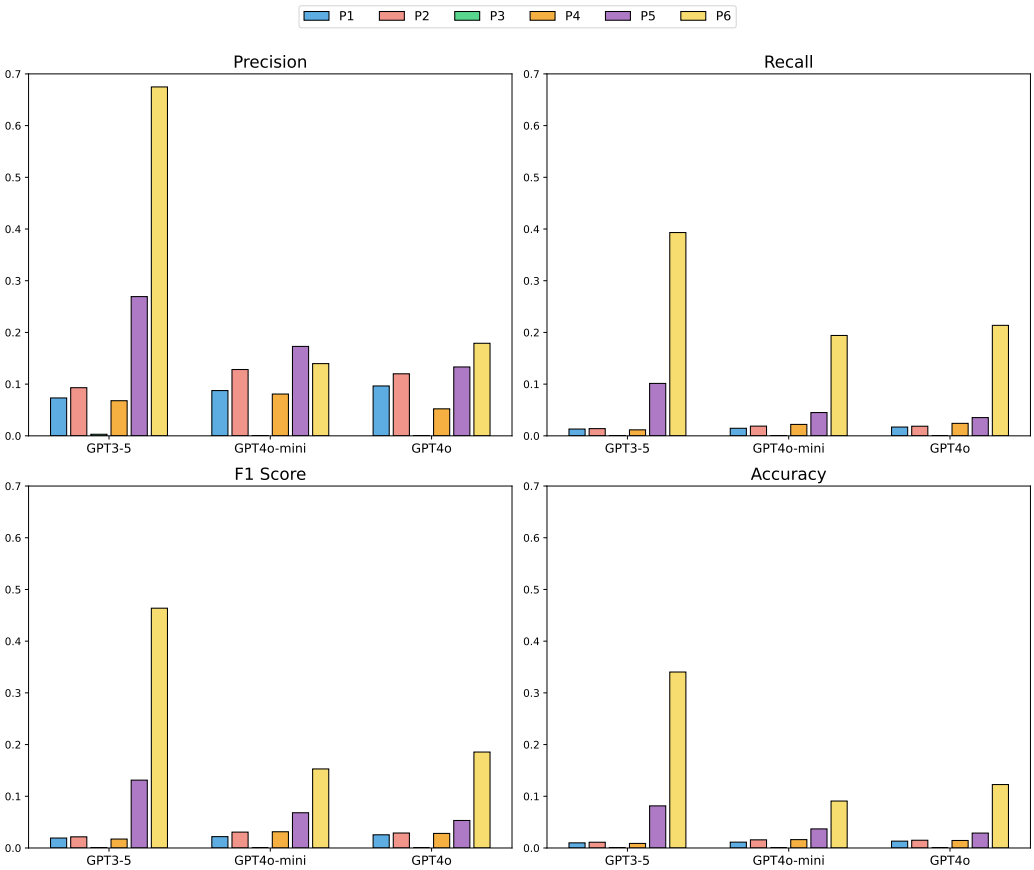


Figure 4. Performance comparison of different LLMs and prompts on the concept generation task [37]. For consistency, the *y*-axis ranges across metrics were standardized to facilitate visual comparison. The *x*-axis denotes different LLMs and prompt configurations (P1–P6), while the *y*-axis denotes the evaluation metric scores (Precision, Recall, F1 Score, and Accuracy, ranging from 0 to 1).

To statistically validate the above trends, we conducted within-course non-parametric tests along two complementary axes. (i) *Cross-model, fixed prompt.* For each prompt, we compared GPT-3.5, GPT-4o-mini, and GPT-4o using a Friedman test (Table 5). Minimal context (P1) yields no significant cross-model differences, whereas modest added context (P2–P4) produces highly significant gaps across metrics (all $p < 0.01$). Under richer inputs (P5 and P6), Precision and F1 remain significantly different across models (P5: $p < 0.01/p < 0.05$; P6: $p < 0.01/p < 0.05$), while Recall differences diminish (often n.s.), suggesting recall saturation once prompts become sufficiently informative. (ii) *Within-model, varying prompts.* For each model, we first ran a Friedman test across P1–P6 and found omnibus differences to be highly significant for Precision and Recall (all $p < 0.01$). To avoid redundancy, we therefore report the post-hoc pairwise Wilcoxon signed-rank tests with Holm correction (Tables 6–8). For GPT-3.5 (Table 6), enriched prompts (P3–P6) significantly outperform minimal prompts (P1–P2) on both Precision and Recall (mostly $p < 0.01$), whereas Zero-Shot instructions without added content (P4) offer limited gains over P1 (n.s.), indicating that GPT-3.5 benefits primarily from substantive context rather than instruction alone. For GPT-4o-mini and GPT-4o (Tables 7 and 8), nearly all transitions from sparse (P1 and P2) to richer prompts (P3–P6) are significant ($p < 0.01$). Among the most informative prompts, Precision gaps are often small or non-significant (e.g., OneShot vs. ALL), while Recall continues to improve, consistent with a pattern of precision saturation

and continued recall gains as more context is injected. Together, these tests confirm that (a) prompt informativeness systematically shapes performance within each model, and (b) cross-model differences emerge and persist once the prompt contains enough signal to be exploited.

Table 5. Friedman test results comparing LLMs under the same prompt configuration. Values indicate chi-square statistics (df = 2) and corresponding p-values. $p < 0.05$ denotes statistically significant differences and $p < 0.01$ indicates highly significant differences.

Prompt	Precision (χ^2, p)	Recall (χ^2, p)	F1 (χ^2, p)	Accuracy (χ^2, p)
P1	$\chi^2(2) = 3.72$, p = n.s.	$\chi^2(2) = 2.31$, p = n.s.	$\chi^2(2) = 2.99$, p = n.s.	$\chi^2(2) = 2.99$, p = n.s.
P2	$\chi^2(2) = 26.70$, $p < 0.01$	$\chi^2(2) = 28.95$, $p < 0.01$	$\chi^2(2) = 27.98$, $p < 0.01$	$\chi^2(2) = 27.98$, $p < 0.01$
P3	$\chi^2(2) = 11.35$, $p < 0.01$	$\chi^2(2) = 11.35$, $p < 0.01$	$\chi^2(2) = 11.35$, $p < 0.01$	$\chi^2(2) = 11.35$, $p < 0.01$
P4	$\chi^2(2) = 22.05$, $p < 0.01$	$\chi^2(2) = 61.93$, $p < 0.01$	$\chi^2(2) = 33.86$, $p < 0.01$	$\chi^2(2) = 33.86$, $p < 0.01$
P5	$\chi^2(2) = 10.17$, $p < 0.01$	$\chi^2(2) = 5.26$, p = n.s.	$\chi^2(2) = 7.57$, $p < 0.05$	$\chi^2(2) = 7.57$, $p < 0.05$
P6	$\chi^2(2) = 10.23$, $p < 0.01$	$\chi^2(2) = 5.62$, p = n.s.	$\chi^2(2) = 8.54$, $p < 0.05$	$\chi^2(2) = 8.54$, $p < 0.05$

Table 6. Pairwise Wilcoxon signed-rank tests (Holm-adjusted p-values) across prompts for GPT-3.5. Each cell shows Precision/Recall. $p < 0.05$ denotes statistically significant differences, $p < 0.01$ indicates highly significant differences, while n.s. means not significant.

	P1	P2	P3	P4	P5	P6
P1	–	$p < 0.01$/n.s.	$p < 0.01$/$p < 0.01$	n.s./n.s.	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$
P2		–	$p < 0.01$/$p < 0.01$	$p < 0.01$/n.s.	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$
P3			–	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$
P4				–	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$
P5					–	$p < 0.01$/$p < 0.01$
P6						–

Table 7. Pairwise Wilcoxon signed-rank tests (Holm-adjusted p-values) across prompts for GPT-4omini. Each cell shows Precision/Recall. $p < 0.05$ denotes statistically significant differences, $p < 0.01$ indicates highly significant differences, while n.s. means not significant.

	P1	P2	P3	P4	P5	P6
P1	–	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$	n.s./$p < 0.01$	$p < 0.01$/$p < 0.01$	$p < 0.05$/$p < 0.01$
P2		–	$p < 0.01$/$p < 0.01$	$p < 0.01$/n.s.	n.s./$p < 0.01$	n.s./$p < 0.01$
P3			–	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$
P4				–	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$
P5					–	n.s./$p < 0.01$
P6						–

Table 8. Pairwise Wilcoxon signed-rank tests (Holm-adjusted p-values) across prompts for GPT-4o. Each cell shows Precision/Recall. $p < 0.05$ denotes statistically significant differences, $p < 0.01$ indicates highly significant differences, while n.s. means not significant.

	P1	P2	P3	P4	P5	P6
P1	–	$p < 0.01$/n.s.	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$	$p < 0.05$/$p < 0.01$	$p < 0.01$/$p < 0.01$
P2		–	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$	n.s./$p < 0.01$	n.s./$p < 0.01$
P3			–	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$	$p < 0.01$/$p < 0.01$
P4				–	$p < 0.01$/$p < 0.05$	$p < 0.01$/$p < 0.01$
P5					–	n.s./$p < 0.01$
P6						–

Interestingly, GPT-3.5 consistently achieves the highest scores across all automated evaluation metrics. However, a closer examination of the generated outputs reveals that this advantage stems not from a universally higher quality of generation, but from a closer lexical alignment with the ground-truth annotations. In contrast, the concepts produced by GPT-4o and GPT-4o-mini, while not achieving similarly high metric scores, often exhibit strong pedagogical relevance and semantic validity. Upon manually reviewing samples from all models, we found that many of the concepts generated by GPT-4o variants are well-grounded in course content, but differ in expression or level of abstraction from the annotated labels. For instance, GPT-4o may generate terms such as "unsupervised pattern discovery" or "hyperplane optimization" instead of the exact ground-truth terms "clustering" or "support vector machines." These concepts are not incorrect or irrelevant—in fact, they may even offer broader or more insightful representations—but their lexical mismatch leads to lower automatic scores.

This inconsistency between evaluation metrics and actual concept quality underscores a key limitation of string-overlap-based evaluation. As observed in prior studies [9,45], large language models are capable of generating semantically meaningful content that deviates from reference annotations without compromising quality. To account for this discrepancy and more accurately assess generation outcomes, we conducted a follow-up human evaluation (Section 5.2) in which domain experts evaluated the quality and relevance of generated concepts beyond literal matching. This qualitative perspective complements the quantitative analysis and provides a more reliable understanding of model performance in open-ended educational settings.

In summary, our ablation study demonstrates that both the granularity of input context and the choice of LLM variant significantly influence concept generation outcomes. While GPT-3.5 excels under current evaluation metrics, GPT-4o produces outputs that are often more abstract or semantically rich, yet undervalued by surface-based scoring. These findings underscore the importance of integrating both quantitative and qualitative evaluations when assessing large language models in educational applications.

5.2. Human Evaluation on Concept Generation

5.2.1. Quantitative Analysis

While metrics-based evaluation method offers a convenient way to compare model outputs, they often fall short in capturing the true quality of generated content, particularly when the generated concepts are semantically appropriate but differ lexically from the annotated ground truth. As discussed in Sections 5.1.1 and 5.1.2, it is important to note that the ground-truth concepts in the MOOCCube dataset were initially generated by a neural model based on course subtitles and subsequently refined through manual annotation. Although human annotators improved the quality and correctness of the extracted concepts, the ground truth remains inherently constrained by the limitations of traditional text-based extraction methods. Specifically, it tends to focus on concepts explicitly mentioned in the text, making it difficult to capture broader, implicit, or abstract concepts that are essential for fully understanding the course content. Consequently, evaluation metrics such as Precision and F1 Score may penalize valid but lexically divergent outputs. To overcome these limitations and obtain a more accurate assessment of concept quality, we conducted a human evaluation involving domain experts.

We recruited four expert annotators, each with strong familiarity in their respective subject areas, to assess the quality of LLM-generated course concepts. Three LLM variants (GPT-3.5, GPT-4o-mini, and GPT-4o) were evaluated across six prompt configurations (P1–P6). For each model–prompt combination, we randomly sampled 20 courses and selected 10 generated concepts per course. In addition, the corresponding ground-truth

concepts were included for reference comparison. Each concept was independently rated on a 5-point Likert scale, with scores reflecting a holistic judgment based on both conceptual correctness and course relevance:

- 1 point: Irrelevant or fundamentally incorrect concept
- 2 points: Marginally relevant or low-quality/incomplete expression
- 3 points: Generally valid, but ambiguous or weakly related to the specific course
- 4 points: High-quality concept that helps understanding of the course content
- 5 points: Core concept that clearly belongs to the course and significantly aids comprehension

Table 9 presents the average human evaluation scores for each model–prompt combination. Several key insights emerge from this evaluation: first, LLM-generated concepts consistently outperform the ground-truth concepts from the MOOCCube dataset across all prompts and models. While the ground truth maintained a fixed average score of 2.677, LLM-generated outputs achieved notably higher scores, reaching up to 3.7. This confirms the hypothesis raised in Sections 5.1.1 and 5.1.2—namely, that metric-based evaluations systematically underestimate LLMs' true performance due to their reliance on surface-level string matching. In contrast, human evaluators were able to identify semantically appropriate and pedagogically valuable concepts, even when those differed lexically from the reference set. The ground truth, generated through neural models trained on subtitles, shares the same limitations as traditional NLP baselines: a dependence on local textual patterns and limited abstraction. The human evaluation thus provides strong validation of LLMs' capacity to infer meaningful concepts beyond the literal text.

Table 9. Average scores of human evaluation for each prompt and model [37].

Model/Prompt	P1	P2	P3	P4	P5	P6
Ground Truth	2.677	2.677	2.677	2.677	2.677	2.677
GPT-3.5	3.613	3.454	**3.630**	3.346	3.083	3.205
GPT4o-mini	3.620	3.461	3.320	3.376	3.341	3.276
GPT4o	**3.700**	**3.516**	3.478	**3.435**	**3.519**	**3.573**

Second, GPT-4o achieved the highest overall scores, outperforming both GPT-4o-mini and GPT-3.5 across nearly all prompt configurations. Its particularly strong performance under P1 (*Zero-Shot*), P2 (*One-Shot*), and P6 (*Subtitle*) highlights two complementary capabilities: robustness in sparse input settings and the ability to effectively process rich contextual data. This dual strength echoes the findings from Section 5.1.2, where GPT-4o demonstrated stable improvements as more information was provided. By contrast, GPT-3.5 performed best under P3 but showed noticeable performance drops under denser prompts like P5, suggesting that excess input complexity or noise may impair its generation quality. These patterns suggest that prompt–model compatibility plays a key role in generation effectiveness, particularly for smaller or less capable models.

Third, the relative performance across prompt types reveals that more context is not always beneficial. Although prompts P5 and P6 contain the most detailed information, including full subtitle transcripts, their scores do not uniformly exceed those of simpler prompts. In fact, P1 and P2, where minimal information is given, often lead to higher scores, especially for GPT-4o. This may seem counterintuitive, but it reflects the fact that LLMs, when given only the course name or brief description, tend to produce broad, high-level concepts that align well with course concepts without introducing noise. In contrast, dense inputs such as subtitles can include irrelevant or overly specific information that dilutes

output quality. This issue is particularly pronounced for GPT-4o-mini and GPT-3.5, which appear more susceptible to information overload.

Fourth, GPT-4o shows relatively consistent performance across all prompts, with small variation in average scores. This suggests a higher degree of generalization capability, allowing it to generate high-quality outputs even when inputs vary significantly in structure and completeness. Its internal representation of educational content appears strong enough to support coherent concept generation under both minimal and maximal contexts. In comparison, GPT-3.5 displays a narrower operating range—it performs well when given structured yet moderate input but struggles under either sparse or overly detailed conditions.

Table 10 reports the inter-rater reliability of human evaluation using Fleiss' κ. The overall agreement across all annotators was 0.09, which falls into the "slight" range according to Landis and Koch [46]. Per-condition values ranged from -0.00 to 0.18, with the ground-truth concepts achieving the highest agreement ($\kappa = 0.178$). These results indicate that, while experts occasionally diverged in their judgments, such variability is not unexpected given the inherently subjective nature of evaluating concept quality. Several factors contributed to these differences. One key factor is that the evaluated courses covered a broad range of disciplines (e.g., computer science, engineering, humanities, and social sciences), making it natural for experts to be more confident in domains closer to their expertise while being more variable in unfamiliar areas. Another factor is that experts held different preferences regarding concept granularity: some favored broader, integrative notions that highlight thematic structures, while others emphasized fine-grained technical terms, leading to discrepancies in scoring. In addition, individual evaluative habits and interpretive styles also introduced variation, particularly when concepts were semantically valid but expressed at different levels of abstraction. Nevertheless, despite this variability, all experts consistently agreed that LLM-generated concepts were pedagogically superior to the ground-truth concepts (as shown in Table 9), underscoring the robustness of our overall findings.

Table 10. Inter-rater reliability of human evaluation: Fleiss' κ for each prompt–model combination and overall agreement. Interpretations follow Landis and Koch [46].

Model/Prompt	P1	P2	P3	P4	P5	P6
GPT-3.5	0.027	0.036	0.035	0.145	0.100	0.115
GPT4o-mini	0.008	0.025	0.023	0.070	0.064	0.102
GPT4o	−0.0003	0.157	0.036	0.076	0.066	0.047
Ground Truth	0.178					
Overall (all conditions)	0.090 (Slight agreement)					

Taken together, these results provide a more nuanced view of model performance and prompt design. They suggest that the best-performing configuration is not necessarily the most information-rich one, and that model scale and architectural differences interact meaningfully with input complexity. These findings reinforce the importance of tailoring prompts to model capacity in real-world educational applications and further demonstrate that human evaluation is indispensable for uncovering generation quality that may be hidden under surface-level metric assessments.

5.2.2. Case Study

To complement the quantitative findings, we conducted a small-scale case study to qualitatively examine the characteristics of concepts generated by different approaches. Specifically, we compared the outputs of (1) traditional NLP baselines such as TF-IDF

and TextRank, (2) the ground-truth annotations in the MOOCCube dataset, and (3) LLM-generated outputs. The goal of this comparison is to explore differences in conceptual granularity, abstraction level, and alignment with instructional content, particularly the extent to which LLMs can go beyond surface extraction to produce pedagogically meaningful and structurally coherent concepts.

As shown in Figure 5, we selected two representative courses to illustrate these contrasts in depth. Across both courses, LLM-generated concepts demonstrate a noticeable improvement in instructional value compared to the other sources. Rather than producing isolated terms, LLMs tend to generate concepts that are thematically cohesive and instructional in tone, often resembling course module titles or learning objectives. For example, in the *Advanced C++ Programming* course, while traditional methods retrieve terms like *Function* or *Pointer*, LLMs output higher-level and more pedagogically framed concepts such as *Object-oriented programming*, *Inheritance and polymorphism*, and *Lambda expressions*. These are not just code-level keywords but reflective of broader programming paradigms that structure how the course content unfolds. Moreover, LLM-generated concepts span different levels of abstraction, from overarching themes down to concrete implementation details. This layering effect is particularly evident in both courses. In the *Mental Health Education for College Students* course, for instance, terms like *Mental health literacy* and *Cognitive-behavioral techniques* appear alongside *Emotion regulation* and *Mindfulness training*, forming a blend of foundational knowledge, psychological models, and applicable coping strategies. This balance is rarely found in concepts extracted by statistical methods or annotated via surface-level heuristics.

Another key distinction lies in the coherence of concept groupings. LLM-generated lists often exhibit internal logical structure, with adjacent terms complementing or expanding upon each other. In contrast, ground-truth and baseline results tend to be either too fragmented or too generic to support instructional scaffolding. While *Stress* or *Belief* are relevant terms, they lack the precision and framing that would make them effective as units of teaching or assessment. Perhaps most notably, some LLMs' outputs go beyond what is explicitly mentioned in the course subtitles. Concepts such as *Smart pointers* or *Cognitive-behavioral techniques* do not always surface in raw textual data but are inferred from broader context. This suggests that LLMs are capable of synthesizing knowledge in a way that mirrors expert-level curriculum reasoning, rather than merely extracting patterns. These examples reinforce the potential of LLM-based models to generate concepts that are not only relevant but also pedagogically aligned, structurally organized, and instructionally versatile. This capacity makes them strong candidates for supporting downstream applications such as syllabus design, automated curriculum modeling, or personalized learning path generation.

5.2.3. Expert Feedback

To enrich the human evaluation with qualitative insights, we conducted follow-up interviews with all four expert annotators. While the Likert-scale scores provided a structured assessment of concept correctness and relevance, the interviews aimed to elicit pedagogical considerations and evaluative dimensions not easily captured through quantitative measures. All experts were provided with course descriptions and a representative subset of concepts in advance to ensure contextual familiarity. Each expert participated in a semi-structured interview lasting approximately 10 min, during which we asked about their overall impressions of LLM-generated concept quality, any instances where LLMs generated unexpectedly high-quality concepts, and their preferences regarding the desired granularity of concepts for instructional purposes. These interviews yielded deeper in-

sights into expert perceptions and highlighted nuanced factors influencing the evaluation of concept suitability and educational effectiveness.

Course: Advanced C++ Programming	
Ground Truth	Abstract data type, Dynamic memory allocation, Exception handling, Linked list, Interface
TF-IDF	Function, Template, Inheritance, Pointer, Constructor
GPT4o	Object-oriented programming, Inheritance and polymorphism, Smart pointers, Exception handling mechanisms, Lambda expressions
Expert Comment	LLM-generated concepts span multiple levels of abstraction, from high-level programming paradigms to specific language features.

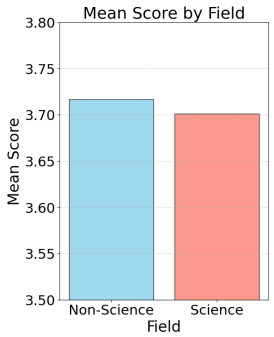
Course: Mental Health Education for College Students	
Ground Truth	Emotional regulation, Self-esteem, Anxiety, Identity, Stress
TF-IDF	Worry, Psychological counseling, Mental state, Belief, Emotional response
GPT4o	Mental health literacy, Emotion regulation, Stress coping strategies, Cognitive-behavioral techniques, Mindfulness training
Expert Comment	LLM-generated concepts are thematically cohesive and pedagogically actionable, reflecting both the conceptual scope and applied skills targeted by mental health education.

Figure 5. Comparison of concepts generated by different methods for two case courses [37].

A recurring theme throughout the interviews was the role of concept granularity in supporting learning. Experts noted that while technical precision is important, concepts that are too fine-grained may overwhelm students, particularly those unfamiliar with the subject matter. Instead, broader, thematically cohesive concepts were considered more effective in introducing course topics and guiding learner attention. This viewpoint aligns with the evaluation patterns observed in Table 9, where generalized concepts often received higher scores than narrowly scoped or overly specialized ones. Beyond this pedagogical observation, the experts expressed a high level of satisfaction with the quality of LLM-generated concepts. Many described the outputs as "surprisingly relevant" and "reflective of actual instructional intent". Some even noted that LLM-generated concepts could serve as valuable input for course syllabus design or formative assessments. Compared to ground-truth concepts or NLP baselines, the LLMs' outputs were frequently praised for their semantic coherence and instructional usefulness.

An interesting disciplinary distinction also emerged from the interviews. According to one expert, LLMs exhibited different tendencies when applied to different domains. In science and engineering courses, the models often generated specific technical terms that aligned with canonical topics. In contrast, for humanities and social science courses, the outputs tended to be more abstract and integrative. This observation prompted a comparison of average human evaluation scores across disciplines. As shown on the left side

of Figure 6, non-science courses received slightly higher scores than science courses. The example concepts on the right side further illustrate this: in *The Historical Career and Methodology*, LLMs generated overarching ideas such as *Development Trends in Historiography*, whereas in *High-Frequency Electronic Circuits*, it produced precise terms like *LC Oscillator* and *High-Frequency Oscillation*. This difference suggests that LLMs' generative strength in abstraction may be particularly well-suited for concept modeling in non-technical domains. These insights highlight the importance of combining expert judgment with quantitative evaluation. They also suggest that LLM-generated concepts, when appropriately interpreted, can meaningfully support educational design across diverse subject areas.



Concept Generation Results

Field	Course	Concepts
Non-Science	The Historical Career and Methodology	Historical Source Analysis Development Trends in Historiography
Science	High-Frequency Electronic Circuits	High-Frequency Oscillation LC Oscillator

Figure 6. Human evaluation of the science and non-science courses, along with examples of concept generation in both fields [37].

5.3. Performance on Concept Extraction

To further evaluate the reasoning capabilities of LLMs, we introduced a constrained concept extraction task, which differs fundamentally from the open-ended nature of the concept generation task. Instead of generating concepts freely, the model is required to identify relevant concepts from a predefined candidate list. This setup reflects a more structured decision-making process and enables us to assess the model's ability to perform fine-grained semantic discrimination under explicit constraints. In designing this task, we adopted two strategies to construct the candidate concept list, each intended to probe different levels of semantic interference. For each target course, the list was composed of (a) its own concepts combined with those from a randomly sampled course in a different domain, or (b) concepts from a course within the same domain. The former setting presents more distinct conceptual boundaries, while the latter increases semantic overlap and therefore the difficulty of discrimination. This setup allows us to assess not only whether LLMs can recognize course-relevant concepts but also how they respond to near-domain distractors, offering a more rigorous test of their reasoning ability. Across all settings, we evaluated three LLM variants and six prompt configurations, the same as in the generation task. As shown in Figure 7, several trends emerge across models and prompt designs: (1) GPT-4o consistently outperforms both GPT-3.5 and GPT-4o-mini across all four evaluation metrics, reinforcing its strength in constrained reasoning tasks. Performance generally improves with the addition of contextual input, with mid- to high-information prompts (P3 to P6) yielding higher accuracy and F1 scores than minimal prompts (P1 and P2). However, the gains from additional context vary by model. For GPT-3.5 and GPT-4o-mini, overly detailed prompts can introduce irrelevant information or semantic noise, leading to marginal or even negative effects on extraction accuracy. (2) We also observe that model performance is sensitive to the composition of the candidate concept list. When the distractor concepts come from a different domain, all models perform more confidently, benefiting from clearer conceptual separability. In contrast, the same-domain setting poses a greater challenge due to increased semantic similarity. Nonetheless, GPT-4o maintains strong performance even under this more difficult condition, suggesting robust semantic understanding and

generalization beyond simple keyword matching. (3) These results reveal a meaningful interaction between model scale and prompt structure. While large-scale models like GPT-4o successfully leverage additional context to refine their predictions, smaller models struggle to integrate dense information, often becoming susceptible to distraction. This suggests that effective performance in extraction tasks is not merely a function of prompt length, but depends on the model's ability to prioritize relevant content within constrained formats.

Overall, our analysis of the concept generation and the concept extraction tasks suggests that LLMs are highly capable of generating and extracting course-related information, potentially reducing the time and effort required by educators.

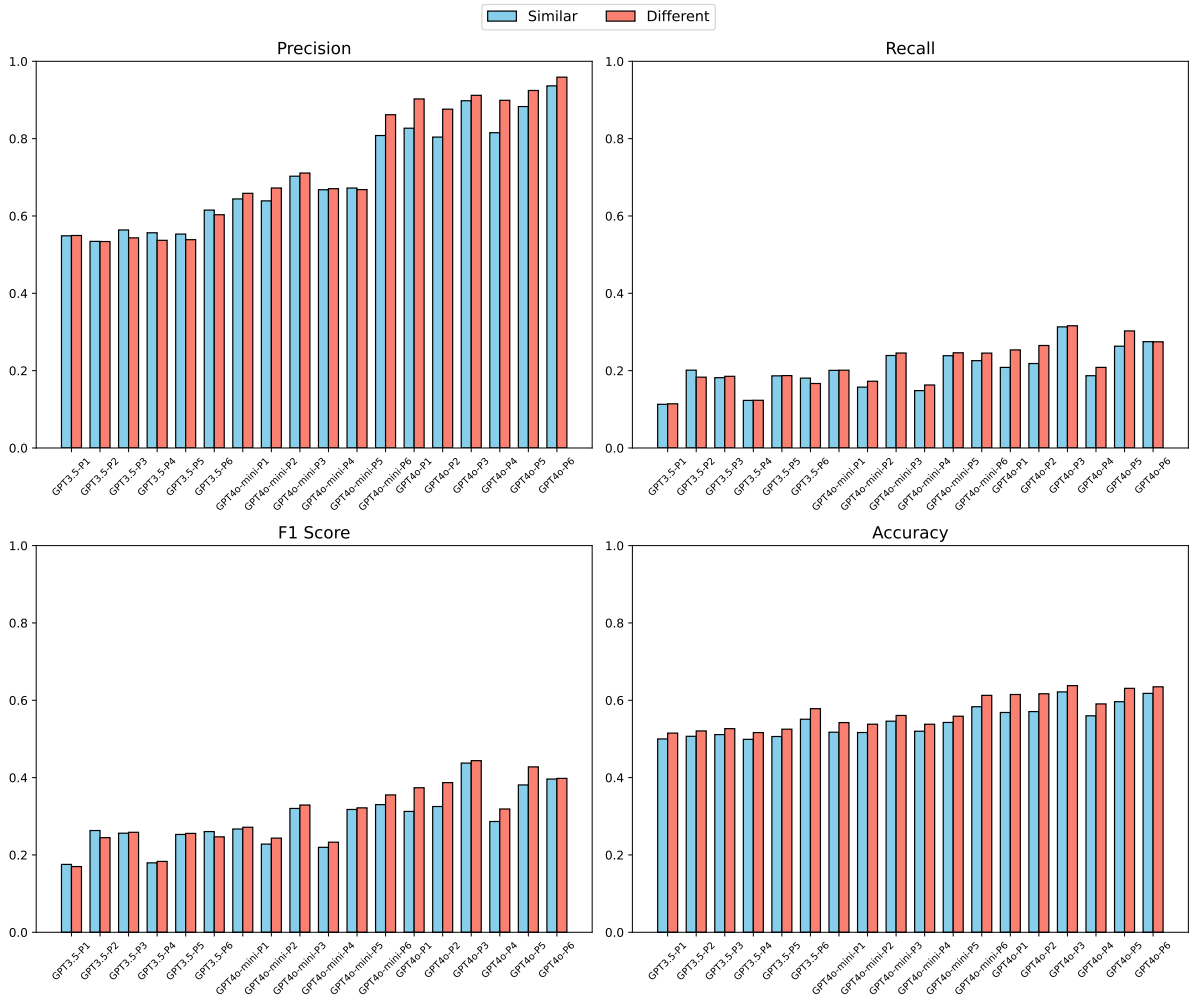


Figure 7. Performance comparison of different LLMs and prompt configurations on the concept extraction task [37]. The y-axis ranges were unified across metrics to enable clearer comparison between models. The x-axis denotes different LLMs and prompt configurations (P1–P6), while the y-axis denotes the evaluation metric scores (Precision, Recall, F1 Score, and Accuracy, ranging from 0 to 1).

5.4. Performance on Relation Identification

Beyond recognizing individual concepts, understanding the prerequisite relationships between them is critical for modeling knowledge structures and designing effective learning trajectories. In this task, we evaluate whether LLMs can infer such inter-conceptual dependencies, which often involve implicit and context-dependent reasoning beyond surface-level matching. Each model was presented with 100 concept pairs and tasked with assigning a scalar score in the range of $[-1, 1]$, indicating the likelihood that one

concept serves as a prerequisite for the other. To systematically assess the impact of input information, we employed six prompt configurations varying in granularity, from minimal descriptions to enriched definitions and course-level context. However, many prerequisite relations are not explicitly stated in course materials, further increasing the task's difficulty. Note that our dataset is restricted to computer science and mathematics courses, and thus does not contain interdisciplinary concept pairs. Consequently, we cannot directly evaluate model robustness on cross-domain relations, although we acknowledge that such settings may pose additional challenges.

As shown in Figure 8, GPT-4o consistently achieves the highest performance across all four evaluation metrics, reflecting its superior ability to reason about inter-concept dependencies. In general, richer prompts (e.g., P5 and P6) lead to improved results, confirming the benefit of contextual input. However, this trend is not uniform across models. For GPT-3.5 and GPT-4o-mini, the performance gains from additional information plateau or even regress, particularly in terms of recall. This suggests that while richer context can aid inference, it may also introduce semantic noise that overwhelms smaller models, reducing their confidence in making relational predictions. In contrast, GPT-4o appears more capable of leveraging complex input while maintaining prediction precision.

Interestingly, we observe that recall performance for GPT-4o slightly drops under the most informative prompt, despite its strong precision. One plausible explanation is that stronger models tend to adopt a more conservative inference style when faced with ambiguous semantic patterns or insufficient causal cues. Rather than over-asserting relations, they default to caution, leading to fewer false positives but also more false negatives.

The intrinsic difficulty of the task was further confirmed through a small-scale human evaluation. Four domain experts were asked to manually annotate the same set of 100 concept pairs, and all reported that determining prerequisite relationships was nontrivial, especially for loosely defined or abstract concepts. To further contextualize model performance, Figure 9 presents three representative cases that were particularly challenging. For clarity, we interpret model predictions using discrete labels: 1 indicates Concept A is a prerequisite of Concept B, -1 indicates the reverse, and 0 denotes no identifiable prerequisite relation. In all three cases, annotators expressed uncertainty or disagreement about the directionality, yet LLMs produced predictions consistent with the ground truth. This suggests the model's ability to capture implicit semantic dependencies that are not always made explicit in instructional materials. The first case, *Multiplication* $\rightarrow$ *Function*, involves foundational mathematical concepts. Although multiplication often underpins the understanding of algebraic functions, the dependency is rarely made explicit in curricula. Experts acknowledged this, and LLMs correctly identified the latent prerequisite relationship. The second case, *Parity* $\rightarrow$ *Integer (Reverse)*, is particularly subtle. While parity depends on the concept of integers, the two are closely linked, and several annotators were unsure about whether a directional prerequisite could be definitively assigned. LLMs' reverse-direction prediction matched the ground truth and reflected a reasonable conceptual interpretation. The third case, *Network Architecture* $\rightarrow$ *Dynamic Memory Allocation*, exemplifies a failure instance. Though the ground truth labels architecture as a prerequisite, the relationship depends heavily on curricular framing. Experts were divided in their annotations, and LLMs defaulted to predicting no dependency. While incorrect, the output reflects the model's cautious behavior under semantic uncertainty. These examples illustrate both the reasoning potential of large language models and the inherent ambiguity of prerequisite relation identification. They further support the view that LLMs' performance in this task, while imperfect, represents meaningful progress toward modeling instructional structures.

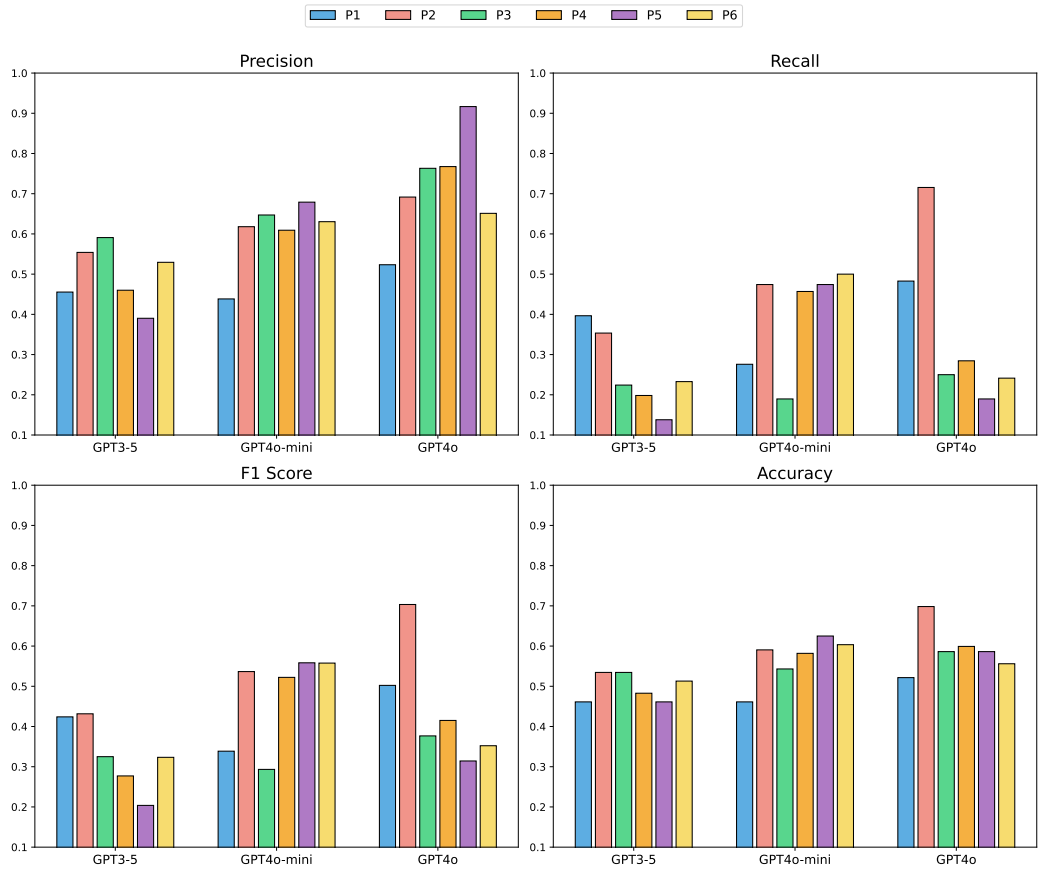


Figure 8. Performance comparison of different LLMs and prompt configurations on inter-conceptual relation identification task [37]. The y-axis ranges were unified across metrics to enable clearer comparison between models. The x-axis denotes different LLMs and prompt configurations (P1–P6), while the y-axis denotes the evaluation metric scores (Precision, Recall, F1 Score, and Accuracy, ranging from 0 to 1).

Figure 10 visualizes the distribution of discretized predictions (-1, 0, $+1$) across all prompt–model configurations. Several systematic trends are evident. GPT-3.5 shows the widest fluctuations: under some prompts (e.g., P6) it produces many reverse (-1) predictions, while under others (e.g., P2–P3) the majority collapse into 0, highlighting its sensitivity to prompt design and relatively unstable reasoning. GPT-4o, in contrast, concentrates strongly on 0 with a selective use of $+1$, rarely outputting -1. This pattern suggests a cautious inference style: the model only asserts a prerequisite when it encounters strong supporting cues, and otherwise defaults to "no relation." Such conservativeness explains GPT-4o's superior precision (Figure 8), as it avoids false positives at the expense of lower recall. GPT-4o-mini behaves differently—it produces more $+1$ predictions and fewer 0s across most prompts, indicating a more assertive inference style that favors recall but risks misclassifying ambiguous pairs as prerequisites. Across all models, reverse predictions (-1) remain sparse. This scarcity reflects an intrinsic asymmetry in the task: even for humans, it is cognitively easier to recognize a forward prerequisite ("A is needed for B") or to judge the absence of a relation than to confidently assert the reverse direction ("B is a prerequisite for A"), which requires more explicit curricular evidence. The fact that LLMs rarely predict -1 therefore mirrors human difficulty and the data distribution itself, where forward dependencies dominate. Taken together, the distributions confirm that the outputs are not random but reveal distinct inference tendencies. GPT-4o prioritizes reliability through cautious prediction, GPT-4o-mini leans toward aggressive identification of forward links, and GPT-3.5 oscillates between neutrality and over-assertion depending

on prompt structure. These behavioral signatures not only validate the methodological design (the models clearly differentiate between output classes) but also delineate the scope of current LLMs: while capable of capturing forward dependencies, they remain challenged by reverse relations and often hedge toward neutrality when explicit signals are lacking.

Success Case ✓: Multiplication → Function

Concept A (Prerequisite): Multiplication
Concept B (Dependent): Function
Ground Truth: 1 **LLM Prediction:** 1
Relation Rationale: Understanding multiplication is fundamental to grasping algebraic functions.
Expert Comment: Experts reported uncertainty when judging this pair, as the prerequisite relation is rarely explicit in course materials.

Success Case ✓: Parity → Integer (Reverse)

Concept A (Prerequisite): Parity
Concept B (Dependent): Integer
Ground Truth: -1 **LLM Prediction:** -1
Relation Rationale: The notion of parity assumes prior knowledge of integers.
Expert Comment: This pair elicited hesitation among experts, who noted that the concepts are conceptually intertwined with no obvious directional dependency.

Failure Case ✗: Network Architecture → Dynamic Memory Allocation

Concept A (Prerequisite): Network Architecture
Concept B (Dependent): Dynamic Memory Allocation
Ground Truth: 1 **LLM Prediction:** 0
Relation Rationale: System-level design knowledge typically precedes memory management implementation.
Expert Comment: Experts were divided on this pair due to its abstraction and curricular dependence. While GPT's output did not match the ground truth, its prediction still reflected a cautious and interpretable decision in a context where even humans struggled.

Figure 9. Representative success and failure cases in the relation identification task [37].

While the output distributions highlight distinct behavioral tendencies across models, a more fine-grained view can be obtained by analyzing how these predictions align with ground truth. The quantitative breakdown in Table 11 reveals that the vast majority of errors (80.2%) stem from *failures to infer implicit relations*. This pattern aligns with the intrinsic challenge of prerequisite identification: many course materials do not state prerequisite links explicitly, requiring models to rely on contextual inference and background knowledge. When such cues are absent or ambiguous, models tend to default to predicting "no relation", resulting in high false negative rates. By contrast, only 19.8% of errors were due to *directionality confusions*, where the model correctly identified a dependency but inverted its direction. Although less frequent, these mistakes are still important because directionality is critical for constructing valid learning paths; a reversed edge can mislead learners about knowledge order. The dominance of implicit-relation failures also resonates with the human evaluation results: even domain experts expressed uncertainty when judging many

pairs, particularly those involving abstract or loosely defined concepts. In addition, the concentration of our dataset in computer science and mathematics exacerbates this difficulty. These fields contain numerous semantically related concepts (e.g., *data structures* vs. *algorithmic complexity*) whose relationships depend heavily on curricular framing, thereby increasing the likelihood of both false negatives and directional confusions. Taken together, these findings suggest that improving prerequisite modeling will require not only stronger language models but also richer instructional context and explicit curricular annotations.

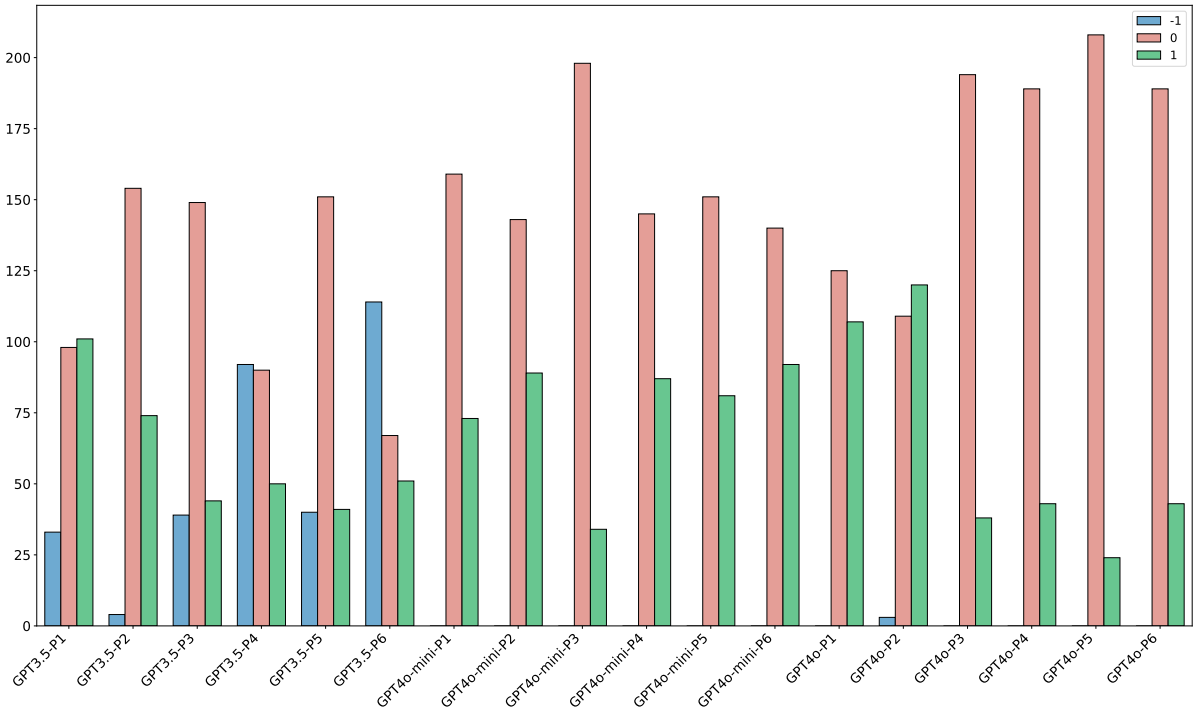


Figure 10. Distribution of LLMs' discretized outputs (-1, 0, $+1$) across all prompt–model combinations. The x-axis denotes different LLMs and prompt configurations (P1–P6), while the y-axis indicates the number of predictions assigned to each discrete label (-1, 0, $+1$).

To further explore the upper bound of model performance on this task, we conducted a preliminary test using GPT-o1-mini, a larger variant beyond our main model set. Although we did not perform a full-scale evaluation due to computational constraints, o1-mini achieved remarkable results on a small sample of 10 concept pairs, yielding perfect precision (1.0), a recall of 0.8, and an F1 score of 0.89. While these results are only indicative, they reinforce the trend that stronger models offer tangible benefits in complex relational reasoning. We leave a more systematic evaluation of o1-mini for future work.

From an educational standpoint, the ability to automatically infer prerequisite relations has significant implications. Such relations form the backbone of concept hierarchies and course progression design. Accurate identification enables applications such as knowledge graph construction, personalized learning path recommendations, and prerequisite-aware curriculum generation. Our findings suggest that GPT-4o, in particular, is approaching a level of relational reasoning that could support these pedagogical applications. Moreover, the observation that prompt structure and model scale interact meaningfully implies that both input design and model choice should be carefully calibrated when deploying language models for fine-grained semantic tasks in educational domains.

Table 11. Distribution of failure types in prerequisite identification task (aggregated over all 18 prompt–model settings; $N = 3318$ errors).

Failure Type	Count	Share (%)
(i) Directionality confusion	656	19.77
(ii) Failure to infer implicit relations	2662	80.23
Total	3318	100.00

6. Discussion and Implication

6.1. Summary of Key Results

In summary, to comprehensively evaluate the effectiveness of LLMs in generating and extracting course concepts as well as identifying their relationships, we conducted a series of experiments across three tasks: concept generation, concept extraction, and relation identification. Each task was designed to assess the distinct capabilities of LLM-based models under varying levels of contextual input. The experiments were conducted using three LLM variants, including GPT-3.5, GPT-4o-mini, and GPT-4o, and compared against representative baselines spanning statistical, embedding-based, and graph-based methods. The performance was quantitatively assessed using four standard metrics: Precision, Recall, F1 Score, and Accuracy. In addition, human evaluations were incorporated to qualitatively assess the relevance and educational value of the generated outputs. The results are organized into three subsections corresponding to each task.

In the concept generation task (Sections 5.1 and 5.2), we examine whether LLMs can produce course concepts that align with ground truth concepts in the MOOCCube dataset. We compare LLM-generated outputs against six baseline methods, including spanning statistical (PMI, TF-IDF, TextRank), graph-based (TPR), and embedding-based (Word2Vec, BERTScore) approaches. These baselines receive the same subtitle inputs used for LLMs, ensuring fair comparison under identical conditions. To further understand how different levels of context influence performance, we design six prompt configurations (P1–P6) by systematically varying the input information provided to LLMs. These range from minimal context (e.g., course name only) to enriched inputs that include course descriptions, existing concepts, and subtitle transcripts. This ablation study is conducted across all three LLM variants to assess model robustness across scales. In addition to automated evaluation using Precision, Recall, F1 Score, and Accuracy, we conduct a human evaluation with four domain experts to assess the quality and relevance of the generated concepts, providing a qualitative complement to the quantitative analysis.

In the concept extraction task (Section 5.3), we test LLMs' ability to identify relevant concepts from a predefined candidate list. Unlike generation, this task constrains LLMs from making selections rather than producing new terms, allowing us to evaluate their reasoning ability under stricter conditions. The candidate lists are constructed by combining concepts from the target course with those from either a course in a different domain (to test semantic separation) or a course in the same domain (to increase conceptual similarity and difficulty). The experiment is conducted using the same three LLMs and six prompt types as in the generation task, enabling direct comparison across tasks and conditions.

Finally, the relation identification task (Section 5.4) focuses on LLMs' capacity to infer prerequisite relationships between pairs of course concepts. We provide the models with concept names, definitions, and related course information, using six different prompt configurations that vary the granularity of injected information. LLMs are asked to assign a directional score between -1 and 1, indicating the presence and strength of a prerequisite relationship. This task allows us to explore whether LLMs can go beyond surface-level

associations and capture asymmetric dependencies between concepts, which are essential for constructing meaningful learning paths and knowledge graphs.

Among the evaluated models, GPT-4o exhibited the most stable and accurate behavior, highlighting the benefits of larger model capacity in tasks involving both open-ended generation and fine-grained reasoning. A closer analysis of each task reveals unique strengths and implications. In the concept generation task, LLMs were able to produce high-quality, structured, and pedagogically aligned concepts that extended beyond the original textual input. These outputs often reflected course-level themes, technical depth, and teaching objectives, which traditional extraction methods failed to capture. In the concept extraction task, LLMs effectively selected course-relevant concepts from noisy candidate lists, indicating strong contextual understanding and semantic discrimination. Notably, GPT-4o demonstrated resilience even when distractor concepts came from the same domain, suggesting that its reasoning extended beyond superficial keyword cues. For relation identification, which involved identifying prerequisite links between concept pairs, all models faced greater difficulty due to the subtle and implicit nature of such relationships. Nevertheless, GPT-4o again led in performance, and qualitative feedback from expert annotators confirmed that even human evaluators found many pairs nontrivial, underscoring the meaningfulness of the task.

Together, these experiments provide a multifaceted view of LLMs' capabilities in educational NLP tasks. By integrating comparisons with baseline methods, model ablation studies, prompt design analysis, and both automatic and human evaluations, we aim to offer a comprehensive understanding of both the strengths and challenges of using LLMs for educational content analysis. These insights serve as the foundation for practical applications and implications, which we discuss in the following section.

6.2. Educational Applications and Insights

These results offer several important implications for real-world educational applications. First, from the perspective of students, the ability of LLMs to generate high-quality, pedagogically aligned course concepts can significantly enhance learning transparency. LLM-generated concepts can serve as concise summaries of course content, assisting students in quickly grasping key knowledge points and making more informed enrollment decisions. Additionally, in personalized learning systems, LLM-based concept extraction can be employed to trace students' understanding by aligning their responses with underlying knowledge components, thereby supporting targeted feedback, misconception detection, and adaptive learning path adjustment. Second, for instructors and curriculum designers, LLMs provide a scalable tool to enrich course metadata, generate concept lists, and construct knowledge graphs without the intensive manual effort traditionally required. Instructors can leverage LLMs' outputs to design diagnostic assessments, organize course modules around conceptual dependencies, and develop more coherent syllabi. Particularly, the relation identification capabilities demonstrated by LLMs enable the automated modeling of prerequisite structures, which are essential for sequencing instructional materials and scaffolding learning activities effectively. Third, from the perspective of MOOC platform operators and educational administrators, the integration of LLM-generated concepts and relations can address critical issues such as data sparsity and metadata incompleteness, especially for new or underdeveloped courses. By automatically enriching course profiles with concept-level representations and prerequisite mappings, LLMs can facilitate more accurate course recommendations, improve curriculum discoverability, and support the construction of dynamic learning pathways. This is especially valuable in large-scale online environments where manual curation is infeasible. Moreover, LLMs offer valuable opportunities for educational technology developers and assessment designers. For learning

management system (LMS) developers, integrating LLM-driven concept generation and relation identification can enable the creation of intelligent content recommendation engines and dynamic curriculum sequencing tools. For assessment specialists, LLMs' ability to extract and organize concepts supports the automated construction of diagnostic tests, adaptive quizzes, and competency-based evaluations.

Furthermore, the three core tasks, which are concept generation, concept extraction, and relation identification, each align naturally with practical instructional scenarios. Concept generation supports course and textbook development by helping educators enumerate and organize key learning objectives and thematic modules. Concept extraction mirrors classroom and assessment contexts such as multiple-choice questions, open-ended response analysis, and concept mapping exercises, where learners must identify and distinguish relevant knowledge elements under varying degrees of ambiguity. Relation identification directly informs the design of coherent curricula and personalized learning paths by uncovering prerequisite relationships that guide the logical progression of content delivery. In particular, LLMs' robust performance in the concept extraction task, even under challenging same-domain distractor settings, highlights its potential to support formative assessment and automated grading systems. Similarly, the ability to generate semantically rich concepts beyond the literal course descriptions demonstrates its value in enhancing content modeling and syllabus design. The capacity to infer implicit prerequisite relations opens opportunities for automated curriculum scaffolding and adaptive course recommendation systems that dynamically adjust to individual learners' prior knowledge. Beyond independent evaluation, concept generation and concept extraction can be combined into a complementary workflow. Concept generation allows LLMs to propose a broad set of candidate concepts, but this set often requires refinement. Concept extraction can serve as a corrective step, identifying the most relevant and accurate items from within the generated range, thereby improving reliability. This interplay reflects realistic use cases. For instance, students may first rely on extraction to identify the main concepts explicitly present in course materials, and then use generation to explore supplementary or prerequisite concepts that enrich their understanding. Conversely, in automated systems, generation may provide a wide coverage of potential knowledge elements, while extraction mechanisms filter and validate these outputs to support robust course modeling and recommendation.

Finally, these findings inform instructional design and prompt engineering practices. Our results indicate that simply increasing the amount of input information does not always lead to better model performance. Moderately informative prompts often outperformed highly detailed ones, suggesting that structuring input with clarity and intentional focus is crucial. This insight is applicable not only to AI-driven educational tools but also to classroom pedagogy, where excessive content can overwhelm learners and obscure critical learning objectives. In summary, LLM-based models, by automating the generation, extraction, and structuring of educational concepts and relationships, offer a practical and effective means to enhance course content modeling, personalize learning experiences, optimize assessment design, and support scalable educational innovations across diverse stakeholders in the modern learning ecosystem.

While promising overall, our study also surfaced several methodological limitations related to the use of LLM-based models. First, although LLMs, particularly GPT-4o, exhibited strong performance, their outputs still require human refinement to ensure pedagogical alignment, conceptual completeness, and contextual appropriateness. Fully automated deployment without expert review remains risky, especially for high-stakes educational applications. Second, model performance was found to be sensitive to prompt complexity and design. While GPT-4o benefited from moderately enriched inputs, smaller models often exhibited degradation when provided with overly detailed or noisy prompts,

indicating that careful and task-specific prompt engineering is essential for optimal results. Third, despite showing reasonable accuracy in relation identification tasks, LLMs tended toward conservative inference strategies, yielding high precision but somewhat lower recall. This suggests that current models are cautious in asserting prerequisite relationships, which may limit their completeness when constructing curriculum knowledge graphs. Finally, differences across disciplines, such as finer granularity and stricter hierarchical structures in scientific domains, indicate that domain-adaptive prompting strategies may further enhance performance.

7. Conclusions

This study presents a comprehensive evaluation of LLM-based models in the tasks of concept generation, concept extraction, and relation identification within educational contexts. Across extensive experiments, including baseline comparisons, ablation studies, and human evaluations, we found that LLMs, particularly GPT-4o, exhibit strong capabilities in generating pedagogically relevant course concepts and accurately identifying inter-conceptual relationships. Unlike traditional methods that are constrained by surface-level features or textual frequency, LLMs are capable of generating semantically meaningful and instructional concepts that are not explicitly mentioned in the input, thereby addressing longstanding gaps in existing methods. Moreover, LLMs demonstrated the ability to reason under constraints, as shown in the concept extraction task, and to infer subtle prerequisite relationships between concept pairs—tasks that even human experts found challenging. These findings affirm the potential of large language models to support automated curriculum modeling, instructional scaffolding, and content recommendation.

While our experiments provide strong evidence for the feasibility of using LLM-based models in educational concept modeling and relation identification, several limitations remain. First, our evaluation was based on expert assessments, which, while authoritative, may not fully reflect students' perspectives or learning challenges. Future work should incorporate user-centered evaluations involving real learners to better assess the practical educational value of LLM-generated outputs. Second, our experiments were limited to courses from the MOOCCube dataset. Although the dataset covers diverse topics, it may not fully represent interdisciplinary or non-traditional course structures. Broader evaluations across varied educational domains and levels are needed to validate generalizability. Third, we primarily evaluated three LLM variants (GPT-3.5, GPT-4o-mini, and GPT-4o). Expanding comparisons to a wider range of large language models, including open-source and multimodal systems, would help benchmark broader capabilities. Finally, while this study focused on pairwise concept and relation modeling, future research should explore more complex structures such as soft prerequisites, cyclical dependencies, and hierarchical knowledge graphs to better align with real-world curricular demands.

Author Contributions: Conceptualization, T.Y. and B.M.; methodology, T.Y. and T.H.; software, C.G.; investigation, T.Y., T.H. and B.M.; formal analysis, T.Y. and B.R.; visualization, B.R.; data curation, C.G.; writing—original draft preparation, T.Y.; writing—review and editing, B.M. and S.K.; supervision, S.K.; project administration, S.K.; funding acquisition, B.M. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JST SPRING, Grant Number JPMJSP2136, JSPS KAKENHI Grant Numbers JP20H00622 and JP24K20903.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study utilized a publicly available dataset, which can be accessed at http://moocdata.cn/. The GPT-generated course concepts in this study are available from the corresponding author upon reasonable request for non-commercial research purposes.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ma, B.; Lu, M.; Taniguchi, Y.; Konomi, S. CourseQ: The impact of visual and interactive course recommendation in university environments. *Res. Pract. Technol. Enhanc. Learn.* **2021**, *16*, 18. [CrossRef]
2. Ma, B.; Yang, T.; Ren, B. A Survey on Explainable Course Recommendation Systems. In *International Conference on Human-Computer Interaction*; Streitz, N.A., Konomi, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2024; pp. 273–287.
3. Pan, L.; Wang, X.; Li, C.; Li, J.; Tang, J. Course concept extraction in moocs via embedding-based graph propagation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 27 November–1 December 2017; pp. 875–884.
4. Lu, M.; Wang, Y.; Yu, J.; Du, Y.; Hou, L.; Li, J. Distantly Supervised Course Concept Extraction in MOOCs with Academic Discipline. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Volume 1, pp. 13044–13059. [CrossRef]
5. Aytekin, M.C.; Saygın, Y. ACE: AI-Assisted Construction of Educational Knowledge Graphs with Prerequisite Relations. *J. Educ. Data Min.* **2024**, *16*, 85–114.
6. Pan, L.; Li, C.; Li, J.; Tang, J. Prerequisite relation learning for concepts in moocs. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1447–1456.
7. Sun, J.; He, Y.; Xu, Y.; Sun, J.; Sun, G. A Learning-path based Supervised Method for Concept Prerequisite Relations Extraction in Educational Data. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA, 21–25 October 2024; pp. 2168–2177.
8. Gupta, P.; Raturi, S.; Venkateswarlu, P. Chatgpt for Designing Course Outlines: A Boon or Bane to Modern Technology. 2023. Available online: http://dx.doi.org/10.2139/ssrn.4386113 (accessed on 31 July 2025).
9. Yang, T.; Ren, B.; Gu, C.; Ma, B.; Konomi, S. Leveraging ChatGPT for Automated Knowledge Concept Generation. In Proceedings of the CELDA2024: International Conference on Cognition and Exploratory Learning in the Digital Age. International Association for Development of the Information Society (IADIS), Zagreb, Croatia, 26–28 October 2024; pp. 75–82.
10. Ehara, Y. Measuring Similarity between Manual Course Concepts and ChatGPT-generated Course Concepts. In Proceedings of the 16th International Conference on Educational Data Mining, Bengaluru, India, 11–14 July 2023; pp. 474–476.
11. Ma, B.; Khan, M.A.Z.; Yang, T.; Polyzou, A.; Konomi, S. Evaluating the Effectiveness of Large Language Models for Course Recommendation Tasks. In Proceedings of the 33rd International Conference on Computers in Education, Chennai, India, 1–5 December 2025.
12. Alexander, K.; Savvidou, C.; Alexander, C. Who wrote this essay? Detecting AI-generated writing in second language education in higher education. *Teach. Engl. Technol.* **2023**, *23*, 25–43. [CrossRef]
13. Perkins, M.; Roe, J.; Postma, D.; McGaughran, J.; Hickerson, D. Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *J. Acad. Ethics* **2024**, *22*, 89–113. [CrossRef]
14. Barany, A.; Nasiar, N.; Porter, C.; Zambrano, A.F.; Andres, A.L.; Bright, D.; Shah, M.; Liu, X.; Gao, S.; Zhang, J.; et al. ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In Proceedings of the International Conference on Artificial Intelligence in Education, Recife, Brazil, 8–12 July 2024; pp. 134–149.
15. Foster, J.M.; Sultan, M.A.; Devaul, H.; Okoye, I.; Sumner, T. Identifying core concepts in educational resources. In Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, Washington, DC, USA, 10–14 June 2012; pp. 35–42.
16. Manrique, R.; Grévisse, C.; Marino, O.; Rothkugel, S. Knowledge graph-based core concept identification in learning resources. In Proceedings of the Joint International Semantic Technology Conference, Awaji, Japan, 26–28 November 2018; pp. 36–51.
17. Changuel, S.; Labroche, N.; Bouchon-Meunier, B. Resources Sequencing Using Automatic Prerequisite–Outcome Annotation. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 1–30. [CrossRef]
18. Yu, J.; Wang, C.; Luo, G.; Hou, L.; Li, J.; Liu, Z.; Tang, J. Course Concept Expansion in MOOCs with External Knowledge and Interactive Game. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4292–4302.
19. Talukdar, P.; Cohen, W. Crowdsourced comprehension: Predicting prerequisite structure in wikipedia. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Montreal, QC, Canada, 7 June 2012; pp. 307–315.
20. Liang, C.; Wu, Z.; Huang, W.; Giles, C.L. Measuring prerequisite relations among concepts. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1668–1674.

21. Manrique, R.; Pereira, B.; Mariño, O. Exploring knowledge graphs for the identification of concept prerequisites. *Smart Learn. Environ.* **2019**, *6*, 21. [CrossRef]
22. Li, I.; Fabbri, A.R.; Tung, R.R.; Radev, D.R. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6674–6681.
23. Zhang, J.; Lan, H.; Yang, X.; Zhang, S.; Song, W.; Peng, Z. Weakly supervised setting for learning concept prerequisite relations using multi-head attention variational graph auto-encoders. *Knowl.-Based Syst.* **2022**, *247*, 108689. [CrossRef]
24. Reales, D.; Manrique, R.; Grévisse, C. Core Concept Identification in Educational Resources via Knowledge Graphs and Large Language Models. *SN Comput. Sci.* **2024**, *5*, 1029. [CrossRef]
25. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
26. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent Abilities of Large Language Models. *arXiv* **2022**, arXiv:2206.07682. [CrossRef]
27. Qiao, S.; Ou, Y.; Zhang, N.; Chen, X.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Chen, H. Reasoning with Language Model Prompting: A Survey. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 5368–5393. [CrossRef]
28. Wei, X.; Cui, X.; Cheng, N.; Wang, X.; Zhang, X.; Huang, S.; Xie, P.; Xu, J.; Chen, Y.; Zhang, M.; et al. Zero-Shot Information Extraction via Chatting with ChatGPT. *arXiv* **2023**, arXiv:2302.10205. [CrossRef]
29. Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Gu, H.; Shen, T.; Qin, C.; Zhu, C.; Zhu, H.; Liu, Q.; et al. A survey on large language models for recommendation. *World Wide Web* **2024**, *27*, 60. [CrossRef]
30. Yang, T.; Ren, B.; Ma, B.; Khan, M.A.Z.; He, T.; Konomi, S. Making Course Recommendation Explainable: A Knowledge Entity-Aware Model Using Deep Learning. In Proceedings of the 17th International Conference on Educational Data Mining, Atlanta, GA, USA, 14–17 July 2024; pp. 658–663. [CrossRef]
31. Bao, K.; Zhang, J.; Zhang, Y.; Wenjie, W.; Feng, F.; He, X. Large language models for recommendation: Progresses and future directions. In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, Beijing, China, 26–29 November 2023; pp. 306–309.
32. Lekan, K.; Zachary, A.P. AI-Augmented Advising: A Comparative Study of ChatGPT-4 and Advisor-Based Major Recommendations. In Proceedings of the NeurIPS Workshop on Generative AI for Education, the Thirty-Seventh Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.
33. Castleman, B.; Turkcan, M.K. Examining the Influence of Varied Levels of Domain Knowledge Base Inclusion in GPT-based Intelligent Tutors. In Proceedings of the 17th International Conference on Educational Data Mining, Atlanta, GE, USA, 14–17 July 2024; pp. 649–657.
34. Lin, J.; Chen, E.; Han, Z.; Gurung, A.; Thomas, D.R.; Tan, W.; Nguyen, N.D.; Koedinger, K.R. How Can I Improve? Using GPT to Highlight the Desired and Undesired Parts of Open-Ended Responses. In Proceedings of the 17th International Conference on Educational Data Mining, Atlanta, GE, USA, 14–17 July 2024; pp. 236–250.
35. Yang, T.; Ren, B.; Ma, B.; He, T.; Gu, C.; Konomi, S. Boosting Course Recommendation Explainability: A Knowledge Entity Aware Model Using Deep Learning. In Proceedings of the 32nd International Conference on Computers in Education, Asia-Pacific Society for Computers in Education, Quezon City, Philippines, 25–29 November 2024; pp. 360–366.
36. Yu, J.; Luo, G.; Xiao, T.; Zhong, Q.; Wang, Y.; Feng, W.; Luo, J.; Wang, C.; Hou, L.; Li, J.; et al. MOOCCube: A Large-Scale Data Repository for NLP Applications in MOOCs. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; pp. 3135–3142. [CrossRef]
37. Yang, T.; Baofeng, R.; Gu, C.; He, T.; Ma, B.; Konomi, S. Examining GPT's Capability to Generate and Map Course Concepts and Their Relationship. *arXiv* **2025**, arXiv:2504.08856.
38. Lai, K.H.; Yang, Z.R.; Lai, P.Y.; Wang, C.D.; Guizani, M.; Chen, M. Knowledge-Aware Explainable Reciprocal Recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024; Volume 38, pp. 8636–8644.
39. Church, K.; Hanks, P. Word association norms, mutual information, and lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.
40. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
41. Mihalcea, R.; Tarau, P. Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004 ; pp. 404–411.
42. Mikolov, T.; Yih, W.t.; Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GE, USA, 10–12 June 2013; pp. 746–751.

43. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
44. Liu, Z.; Huang, W.; Zheng, Y.; Sun, M. Automatic keyphrase extraction via topic decomposition. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 366–376.
45. Liu, J.; Liu, C.; Zhou, P.; Lv, R.; Zhou, K.; Zhang, Y. Is chatgpt a good recommender? A preliminary study. *arXiv* **2023**, arXiv:2304.10149. [CrossRef]
46. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]