

Automatic Knowledge Component Extraction to Bridge High School Knowledge Gaps for Engineering Students

Mario Antonio Pessoa Santos¹, Mariana Nassif¹, Tyagi Lima¹,
Filipe Dwan Pereira², André Nascimento³, Luiz Rodrigues⁴, Cleon Xavier⁵,
Newarney Torrezão Costa⁵, Gabriel Alves³, Rafael Ferreira Mello³

¹ Cogna Educação S.A., SP, Brasil

² Universidade Federal de Roraima (UFRR),

³ Universidade Federal Rural de Pernambuco (UFRPE)

⁴ Universidade Tecnológica Federal do Paraná (UTFPR)

⁵ Instituto Federal Goiano (IFGoiano)

luizrodrigues@utfpr.edu.br

Abstract. *Many students entering higher education have knowledge gaps from secondary education that contribute to low retention and high dropout rates. This short paper presents a case study that applies Knowledge Component (KC) extraction and recommendation techniques to help first-year engineering students bridge these gaps. Using embedding-based similarity methods, we mapped K-12 mathematical competencies to higher education prerequisites and delivered personalized preparatory resources through short video lectures and AI-generated summaries. Results indicate that the similarity-based approach achieved high validation rates in STEM-oriented disciplines, reliably identifying prerequisite KCs, while performance was less consistent in non-technical courses.*

1. Introduction

In countries marked by high social inequality, access to higher education plays an important role in social mobility [Schwartzman 2004]. However, many students enter university after extended periods away from formal schooling, often balancing academic responsibilities with work, financial challenges, and family duties. These conditions create learning gaps that hinder students' academic progress, a situation that is particularly critical in engineering programs, which are known for higher dropout rates [Feinmann and Hsu Rocha 2024].

One promising strategy to address these learning gaps is to recommend high-quality learning materials tailored to students' profiles. However, the effectiveness of such recommendation systems in education depends on their ability to align instructional resources with specific skills or knowledge components (KC) that students need to recover [Abdelrahman et al. 2023]. In that sense, accurate extraction of KC ensures that the recommended activities are both pedagogically relevant and aligned with the learner's evolving state of knowledge. Without reliable identification of KC, recommendations risk being superficial or misaligned, thus reducing their potential impact on learning outcomes [Pardos and Heffernan 2010].

Recent research has advanced diverse methods for the extraction and tracking of KC, ranging from deep learning approaches such as self-attentive networks and incremental knowledge tracking models [Wong and Ramasamy 2024] to the application of large language models (LLMs) for automated labeling and KC recommendation [Scarlatos et al. 2025]. Despite these advances, most of the contributions remain at the level of controlled experiments or small-scale deployments. In contrast, few works have rigorously evaluated KC-based recommendations in scalable real-world higher education contexts, leaving a gap between methodological innovation and applied impact.

Therefore, this short paper addresses this gap by extracting KC for first-year engineering students in higher education and applying it to a scalable recommendation system. The contribution proposed in this study is two-fold: (i) to evaluate the practical impact of KC-based recommendations on a real-world deployment, thus providing empirical evidence of the role of KC extraction in supporting student’ academic success; (ii) to investigate how these extracted KC align with the competencies of the K-12 curricula, thus offering a systematic approach to connect different levels of education and enabling more coherent learning pathways throughout the school–university transition.

2. Background

2.1. Knowledge Component (KC)

The notion of KCs is an important element in the domains of intelligent tutoring systems and personalized learning. KCs are defined as distinct units of knowledge or skill that learners must acquire and coordinate to successfully perform a task [Koedinger et al. 2012]. They can encapsulate a wide range of cognitive elements, such as concepts, principles, facts, or procedures, and serve as the building blocks for modeling student learning and performance. By decomposing complex tasks into finer-grained components, KCs enable a more precise diagnosis of student understanding and support adaptive instructional strategies [Pavlik Jr et al. 2009].

Automated extraction of KCs from learning materials has attracted increasing attention, as it reduces the effort required for manual annotation, enables scalable personalization in education, and provides mechanisms for systematically tracking student progress [Rivers et al. 2016, Shi et al. 2023]. Much of this research is based on the Knowledge-Learning-Instruction (KLI) framework [Koedinger et al. 2012], which conceptualizes assessment events (e.g. exams, discussions, problem-solving activities) as observable proxies for inferring and positioning a student’s underlying knowledge state.

Despite these benefits, identifying KCs remains a challenge. Manually defining and validating the components often requires extensive domain expertise, is time-consuming, and can lead to inconsistencies between courses or instructors [Abdelrahman et al. 2023]. These limitations highlight the importance of developing automated approaches and data-driven algorithms capable of discovering KCs at scale [Shen et al. 2024].

2.2. Automatic detection of Knowledge Component

Recent advances in Natural Language Processing have increased interest in automatically detecting KCs, given the cost and inconsistency of manual labeling [Shen et al. 2024].

Machine learning methods are now widely used to extract and map skills from educational data, supporting scalable personalization and detailed tracking of student knowledge. For example, [Sampaio de Alencar et al. 2025] proposed a hybrid method that combines automated code analysis with instructor expertise to create more interpretable and pedagogically meaningful KCs in introductory Java courses. Research also shows that both commercial and open-source LLMs can generate KCs that are comparable to expert annotations and more consistent in programming assignments [Fan et al. 2025]. In addition, [Shi et al. 2023] proposes a data-driven approach to automatically discover KCs from student code submissions in a programming course. These results demonstrate that the method can reveal interpretable programming skills and common misconceptions.

Furthermore, previous research highlights the importance of both curriculum structure and high-quality KC representations to understand learners. At the curriculum level, [Stavrinides and Zuev 2023] models course-level knowledge and prerequisite relationships, showing how curriculum design affects student performance and providing guidance on the sequencing of KCs to support effective learning pathways. Similarly, advances in knowledge tracing (KT), from early Bayesian models [Pardos and Heffernan 2010] to recent neural and self-supervised approaches such as adaptive attention networks [Wong and Ramasamy 2024], consistently demonstrate that the quality of KC representations is a key driver of KT performance.

Moreover, recent research points to the convergence among KC extraction, KT, and generative models. [Scarlatos et al. 2025] examined how LLM-processed tutor-student dialogs can reveal latent KCs to improve KT, while [Yang et al. 2025] proposed a GLM-4-based method to automatically construct content graphs. These studies illustrate the growing integration of LLMs into the structuring of domain knowledge. Complementarily, graph-based approaches [Zhang et al. 2023] show how fine-grained learning dependencies can be represented, enriching automated KC detection by contextualizing skills within broader learning trajectories.

Although the benefits of KC automation are recognized, most studies focus on the discovery of KC within a single course or domain [Pavlik Jr et al. 2009, Shen et al. 2024]. This leads to a limited understanding of how KCs are connected across educational levels, which can result in a misalignment between STEM requirements in secondary and undergraduate school [Stavrinides and Zuev 2023]. Additionally, research on the large-scale deployment of KC-based methods is scarce, as existing work mainly explores controlled or small-scale settings [Rivers et al. 2016, Shi et al. 2023, Scarlatos et al. 2025]. Addressing these gaps is critical for developing practical, scalable, personalized learning systems that bridge curricular boundaries.

2.3. Research Questions

Based on the previous context, the objective of this short article is to investigate how the relationship between undergraduate STEM courses and KC of K-12 can be systematically analyzed to support knowledge transfer across educational levels. Thus, our first research question (RQ) is:

RESEARCH QUESTION 1 (RQ1):

To what extent do similarity methods for extracting KC from different levels of education produce reliable and interpretable mappings?

Second, there is a lack of research on the use of KCs at scale. This gap is critical because scalable deployment introduces new challenges, including heterogeneity of data, computational efficiency, and the need to ensure pedagogical relevance across diverse populations of learners. As such, our second RQ is as follows:

RESEARCH QUESTION 2 (RQ2):

To what extent can KC-based extraction and recommendation systems be effectively deployed at scale in real-world educational environments?

3. Method

3.1. Context

This study was conducted over a full academic semester with 3,452 first-year students enrolled in Electrical, Production, Civil, and Mechanical Engineering programs. These programs were selected for their high enrollment rates and repeated reports of dropouts, often attributed to difficulties in consolidating prerequisite mathematical competencies. During the observed period, the cumulative dropout rate reached 45%, motivating the implementation of an adaptive preparatory support system.

At the beginning of the programs, students were invited to complete a diagnostic mathematics assessment designed to identify knowledge gaps in key K–12 domains required for engineering curricula. A total of 242 students completed two assessments designed by our learning experts (each comprising 15 questions) and constituted the sample for subsequent analysis.

Diagnostic evaluation revealed knowledge gaps concentrated mainly in the domains of algebra and geometry. The most prevalent deficiencies included: i) Primitive notions of planar geometry (66%); ii) First-degree equations with one unknown (61%); iii) Equivalence of planar figures (59%); iv) Relative positions of a point and a circumference (56%) v) Addition and subtraction of arcs (54%).

3.2. K-12 Knowledge Components Automatic Extraction

To operationalize the process of annotation of KC, three types of material were encoded as embedding representations and organized within a shared vector space: i) resources for the higher education discipline (texts, transcripts and questions); ii) preparatory lecture content; and iii) K-12 KC identified through diagnostic assessments.

KC extraction and alignment were performed using embedding representations generated with the *OpenAI text-embedding-3-large* model [Adachi et al. 2025], followed by similarity-based retrieval using FAISS [Ghadekar et al. 2023]. This approach enables efficient identification of semantically related learning components at scale, supporting personalized learning support [Shen et al. 2024, Wong and Ramasamy 2024].

3.3. KC Validation

After the embedding and similarity search stages, four domain experts validated the candidate prerequisite mappings. Professors reviewed AI-generated associations between higher education content and K–12 knowledge components, with access to the corresponding course materials. Similarity scores were used to prioritize candidate mappings, while final decisions on prerequisite validity were made exclusively by human experts.

This hybrid approach combines scalable automated retrieval with pedagogical judgment, ensuring that the prerequisite mappings were computationally efficient and educationally sound [Koedinger et al. 2012, Rivers et al. 2016].

3.4. Recommendation Mechanism

The recommendation process integrates the learning gaps identified through diagnostic assessments [Abdelrahman et al. 2023], the extracted KCs for each discipline [Stavrinides and Zuev 2023, Sampaio de Alencar et al. 2025], and a repository of preparatory lectures tagged with the corresponding KCs [Zhang et al. 2023]. Remedial content was delivered through short video lectures sourced from the Stoodi platform¹, a Brazilian educational platform that provides short, focused video lessons primarily designed for the preparation of university entrance exams.

3.5. Data Analysis

To address RQ1, we analyzed academic validation of the automatically extracted KCs. For each course, we computed descriptive statistics that capture: (i) the number of unique KCs identified by the embedding-based retrieval process; (ii) the total number of prerequisite associations proposed; and (iii) the number of associations approved by domain experts. To address RQ2, we examine the participation of students' in the recommended preparatory resources and its relationship with learning outcomes. Platform logs were used to extract interaction metrics, including the number of resources accessed per student and the total time spent viewing the recommended materials. The learning outcomes were operationalized as performance on the initial diagnostic assessment and a final post-intervention test covering the same mathematical domains. We computed aggregate descriptive statistics (means) for all variables to characterize overall trends in usage and achievement.

4. Results

To answer RQ1, table 1 presents the results of the academic validation of the extracted KCs. STEM-oriented courses, including Differential and Integral Calculus II, Engineering Science and Technology, and Projective Technical Drawing, achieved high validation rates. In contrast, courses in the humanities and applied social sciences showed low acceptance of the KC recommendations.

Table 2 presents the results for RQ2. The analysis of student participation in recommended preparatory resources reveals a general trend of academic progress, evidenced by an increase in average test scores from 13 in the initial diagnostic to 21.83 in the final assessment, supported by an average of 4.33 resources accessed and approximately 75 minutes of engagement time per student. Although most of the students showed improvement, extensive participation did not always guarantee performance gains, and, conversely, significant progress was observed despite limited interaction, suggesting that factors beyond the frequency and duration of simple access play a critical role in learning outcomes.

¹<https://stoodi.com.br/>

Tabela 1. Academic Validation Results by Course

Course	Academic Validation of KCs		
	Unique	Total	Approved
Administration and Economics for Engineers	21	40	1
Brazilian Society and Citizenship	7	14	0
Clean Technologies and Waste Treatment	3	4	0
Design Thinking and Innovation of Business Models	5	7	0
Differential and Integral Calculus II	22	36	36
Engineering Science and Technology	3	4	4
Projective Technical Drawing	18	27	25

Tabela 2. Statistics of students' interaction with recommended materials

Metric	Average (STD)	Max	Min
Score on the initial diagnostic test	13 (4.68)	15	1
Score on final test	21.83 (7.59)	29	5
Number of resources accessed per student	4.33 (3.07)	16	2
Time spent on resources (seconds)	4,486.16 (4,125.34)	16,350.00	208.00

5. Discussion

Our results show that embedding-based similarity measures reliably identified prerequisite KCs in STEM fields, with high validation rates in calculus and technical drawing. This aligns with previous findings that automated approaches can uncover interpretable and pedagogically relevant skills, particularly when applied to structured domains like programming or mathematics [Shi et al. 2023]. Furthermore, the distinction between technical and non-technical courses reflects observations that KC extraction benefits from domains with clearly defined cognitive hierarchies. However, our findings also echo concerns raised in previous work that embedding-based methods alone may be less effective in capturing contextual or interdisciplinary knowledge [Shen et al. 2024], as seen in humanities-oriented courses.

In addition, we demonstrated that the KC-based recommendation system successfully mapped these gaps to targeted preparatory resources, with evidence of improvements in the final scores of the students. These findings provide empirical support for the role of KC-driven approaches in bridging educational transitions, while also highlighting the complexity of aligning recommendations with heterogeneous learner profiles [Shi et al. 2023, Fan et al. 2025, Rivers et al. 2016, Pardos and Heffernan 2010].

Moreover, the findings of this study highlight the feasibility of linking K–12 competencies to higher education entry requirements, thus supporting the design of more coherent learning pathways and helping reduce dropout rates in engineering courses. This aligns with the larger goal of leveraging artificial intelligence to address educational inequalities and promote stronger curricular continuity [Schwartzman 2004, Feinmann and Hsu Rocha 2024].

6. Limitations and Future Directions

As a short paper, this work presents a focused proof-of-concept, leading to limitations that must be acknowledged. First, participation in diagnostic assessments with the recommen-

dation system was voluntary, which could introduce selection bias, as students who are more motivated or more aware of their difficulties may be overrepresented. Second, while the embedding-based similarity method showed strong reliability in STEM disciplines, its limited effectiveness in humanities and interdisciplinary courses highlights the risk of domain dependency. Furthermore, the evaluation relied heavily on short-term performance improvements measured through diagnostic tests, without capturing long-term retention or knowledge transfer. These constraints limit the generalizability of the findings.

Future research should expand the scope of KC-based recommendation systems in several directions. An important step is to conduct longitudinal studies that track the sustained impact of recommendations on student success, including retention and progression throughout the semester. Additionally, refining methods for mapping K–12 competencies to higher education requirements in non-STEM domains remains a critical challenge, requiring hybrid approaches that combine algorithmic similarity with expert-driven pedagogical insights. Finally, future work should examine equity-related aspects by testing how such systems can reduce educational disparities between diverse student populations and ensure that adaptive interventions benefit those most at risk of academic exclusion.

Statement on the Use of Artificial Intelligence

Artificial intelligence tools, including *ChatGPT* and *Gemini*, were used to support the development of this manuscript, including brainstorming ideas, structuring the narrative, drafting and revising textual content, improving language clarity and presentation, and assisting with proofreading and formatting. All final decisions regarding content, interpretation, and conclusions remained under the authors' responsibility.

Ethical Concerns

This study uses secondary data within the field of Learning Analytics from Cogna Educational S.A. to support evidence-based institutional management. Data processing was conducted in strict compliance with the organization's information security policies, ensuring data anonymization and full adherence to applicable data protection regulations and ethical standards.

Artifacts Availability

This study's artifacts will be made available from the corresponding authors upon reasonable request and compliance to data protection regulations and ethical standards.

Referências

- Abdelrahman, G., Wang, Q., and Nunes, B. (2023). Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11):1–37.
- Adachi, Y., Sakai, A., and Uehara, M. (2025). Cross-language semantic search using openai embedding. In *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, pages 1–6. IEEE.
- Fan, J., Mihaylova, T., Akram, B., Norouzi, N., Brusilovsky, P., Hellas, A., and Leinonen, J. (2025). Adaptive learning curve analytics with llm-kc identifiers for knowledge

- component refinement. In *Proceedings of the 2025 Conference on UK and Ireland Computing Education Research*, pages 1–7.
- Feinmann, J. and Hsu Rocha, R. (2024). Social mobility and higher education in brazil. *Available at SSRN 4702937*.
- Ghadekar, P. P., Mohite, S., More, O., Patil, P., Mangrulkar, S., et al. (2023). Sentence meaning similarity detector using faiss. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–6. IEEE.
- Koedinger, K. R., Corbett, A. T., and Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798.
- Pardos, Z. A. and Heffernan, N. T. (2010). Modeling individualization in a bayesian networks implementation of knowledge tracing. In *International conference on user modeling, adaptation, and personalization*, pages 255–266. Springer.
- Pavlik Jr, P. I., Cen, H., and Koedinger, K. R. (2009). Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. *Online Submission*.
- Rivers, K., Harpstead, E., and Koedinger, K. (2016). Learning curve analysis for programming: Which concepts do students struggle with? In *Proceedings of the 2016 ACM Conference on International Computing Education Research, ICER '16*, page 143–151, New York, NY, USA. Association for Computing Machinery.
- Sampaio de Alencar, R., Demirtas, M. A., Saha, A. S., Shi, Y., and Brusilovsky, P. (2025). Integrating expert knowledge with automated knowledge component extraction for student modeling. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 307–312.
- Scarlato, A., Baker, R. S., and Lan, A. (2025). Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 249–259.
- Schwartzman, S. (2004). Equity, quality and relevance in higher education in brazil. *Anais da Academia Brasileira de Ciências*, 76:173–188.
- Shen, S., Liu, Q., Huang, Z., Zheng, Y., Yin, M., Wang, M., and Chen, E. (2024). A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, 17:1858–1879.
- Shi, Y., Schmucker, R., Chi, M., Barnes, T., and Price, T. (2023). Kc-finder: Automated knowledge component discovery for programming problems. *International Educational Data Mining Society*.
- Stavrinides, P. and Zuev, K. M. (2023). Course-prerequisite networks for analyzing and understanding academic curricula. *Applied Network Science*, 8(1):19.
- Wong, C. S. Y. and Ramasamy, S. (2024). Architectural adaptation and regularization of attention networks for incremental knowledge tracing. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 757–762.
- Yang, M., Diao, M., Luo, J., Shen, W., and Zhang, C. (2025). Glm-4 based method for automatic construction of content graph. *IEEE Access*.

Zhang, X., Liu, S., and Wang, H. (2023). Personalized learning path recommendation for e-learning based on knowledge graph and graph convolutional network. *International journal of software engineering and knowledge engineering*, 33(01):109–131.