From Course to Skill: Evaluating Large Language Model Performance in Curricular Analytics

Zhen Xu, Xinjin Li, Yingqi Huan, Veronica Minaya,
and Renzhe Yu(✉)

Columbia University, New York, NY, USA
{zx2393,xl3319,yh3755,vminaya,renzheyu}@tc.columbia.edu

Abstract. Curricular analytics (CA) – systematic analysis of curricula data to inform program and course refinement – becomes an increasingly valuable tool to help institutions align academic offerings with evolving societal and economic demands. Large language models (LLMs) are promising for handling large-scale, unstructured curriculum data, but it remains uncertain how reliably LLMs can perform CA tasks. In this paper, we systematically evaluate four text alignment strategies based on LLMs or traditional NLP methods for skill extraction, a core task in CA. Using a stratified sample of 400 curriculum documents of different types and a human-LLM collaborative evaluation framework, we find that retrieval-augmented generation (RAG) is the top-performing strategy across all types of curriculum documents, while zero-shot prompting performs worse than traditional NLP methods in most cases. Our findings highlight the promise of LLMs in analyzing brief and abstract curriculum documents, but also reveal that their performance can vary significantly depending on model selection and prompting strategies. This underscores the importance of carefully evaluating the performance of LLM-based strategies before large-scale deployment.

Keywords: Curricular Analytics · Skill Extraction · Large Language Models · Text Alignment · Higher Education

1 Background

Curriculum is a core component of higher education, shaping students' intellectual growth and preparing students for the workforce, while also serving as a benchmark for program quality and institutional reputation. Given the rapid advancements in digital technology and the digital economy, institutions and educational stakeholders are increasingly seeking automated ways to analyze curriculum documents and generate evidence-based insights for improving curriculum design and delivery [4,8,19]. In this context, curricular analytics (CA) has emerged as a subfield of learning analytics (LA), aimed at facilitating data-driven decision making and improvement in courses and programs [8].

A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15882, pp. 203–211, 2025.
https://doi.org/10.1007/978-3-031-98465-5_26

Despite its promise, CA remains relatively underdeveloped [24], due not only to a historical lack of digital curricula data but also to technical challenges of analyzing texts in a scalable and reliable manner. Curricula documents, such as course catalogs, syllabi, and reading materials, vary widely in structure, granularity, and language use, which makes automation difficult. Recent advances in natural language processing (NLP) have helped CA progress from rule-based to more sophisticated semantic approaches [6,7,10–12,18,21], but major challenges still exist, including the difficulty in extracting fine-grained curricular constructs, the lack of standardized curricular ontologies, and the need for pedagogically grounded reasoning. With these challenges, achieving automated extraction of meaningful insights from curricula is still a considerable hurdle.

Recent advances in large language models (LLMs) offer new possibilities for curricular analytics. Their ability to efficiently extract the semantics of natural language could improve how we analyze and interpret curricular content and help identify complex educational constructs and curricular elements that were previously hard to capture. In addition, their natural language interfaces lower technical barriers, making CA more accessible to educators and researchers without extensive technical expertise. As a result, increasing efforts have been made to incorporate LLMs as analytical tools in CA research [6,13,14,16–18,25]. While these studies have shown some promise of LLM-assisted CA, the reliability and generalizability of this promise across different curricular contexts are still not well understood.

In this study, we systematically evaluate the performance of LLMs versus traditional NLP methods in the context of skill extraction, a core CA task that assesses how well course content aligns with workforce demands through the lens of skills. Skills are essential components of jobs and play a key role in shaping individuals' career outcomes in the labor market. Therefore, systematically examining skills and how they are developed through education is crucial for understanding students' future career trajectories and broader workforce trends [5,23]. By conducting this evaluation, our contributions are twofold. First, we provide one of the first systematic empirical assessments of LLMs' capabilities in curricular analytics, benchmarking their performance against major traditional NLP paradigms commonly used in CA. Second, we examine how LLM performance varies across different prompting strategies, model selections, and curriculum document types, offering practical guidelines and important considerations for researchers and practitioners seeking to integrate LLMs into CA research and applications.

2 Data and Methods

2.1 Datasets

Curriculum Documents. The curriculum documents used in this study come from two sources: (1) Course Syllabi from Open Syllabus[1], a nonprofit archive of over 20.9 million higher education syllabi worldwide; (2) General catalog with

[1] https://www.opensyllabus.org/.

short descriptions of individual courses from a large, urban, public two-year college in the United States, which is publicly available on its official website. We restrict our analysis to courses from the 2017-18 academic year for consistency and use stratified sampling to select 100 curriculum documents from each source, covering a diverse range of major areas and document length categories.

From stratified sampling, we generate four types of curriculum documents commonly used in CA: (1) course descriptions from the general catalog, (2) course descriptions in syllabi, (3) learning outcomes in syllabi, and (4) the combination of course descriptions and learning outcomes in syllabi. More details about the dataset can be found in the Supplemental Information.

Skill Framework. *O*NET* (Occupational Information Network) is a comprehensive database of job characteristics and worker skills from the US Department of Labor, which has been widely used for labor market analysis, curriculum design, and career guidance [1,3,9,10]. We use the *Detailed Work Activity (DWA)* taxonomy from *O*NET*, which includes 2,070 short descriptions of real-world work activities across various occupations, and treat DWAs as skills in our analyses.

2.2 Skill Extraction

A course can cultivate multiple skills, and here we extract the top 10 most relevant skills from each curriculum document for consistency of comparison. Skill relevance is measured by semantic alignment between a skill and a course. We apply the following four text alignment strategies.

Token-Based (TF-IDF): Following [15], we calculate alignment scores between each course and DWA skill using TF-IDF weights, combined with relevance weights based on token importance in the DWA dataset relative to Wikipedia. The weighted TF-IDF scores are summed and normalized by token count to adjust for course length.

Embedding-Based (BERT): We adapt the embedding-based matching method from [10]. SBERT, a siamese network-based model known for effective sentence embeddings [20], is used to calculate alignment scores between each curriculum document and skill description via cosine similarity.

Zero-Shot Prompting (ZERO-SHOT): We use both open-source and proprietary models, including GPT-4o, Llama 3.3-70B, Gemini 1.5 Pro, and Claude 3.5 Sonnet to perform zero-shot skill extraction. Each model is prompted with the curriculum document and the predefined skill description list to perform skill extraction. The full prompt is provided in the Supplementary Information.

Retrieval-Augmented Generative (RAG): We use RAG, a strategy that improves performance by retrieving relevant external information before generation [25], as a pre-filtering step to narrow the skill pool. Specifically, we embed 2,070 skill descriptions into a vector database and retrieve the top 20 most relevant skills[2] based on cosine similarity with the curriculum document. These retrieved skills, along with the curriculum document and query, are then used to construct a structured prompt for the LLM to extract skills.

2.3 Performance Evaluation

To evaluate the performance of each text alignment strategy, We score the alignment of extracted skills on a 5-level scale using a human-LLM collaborative evaluation framework, and aggregate the scores to assess overall performance.

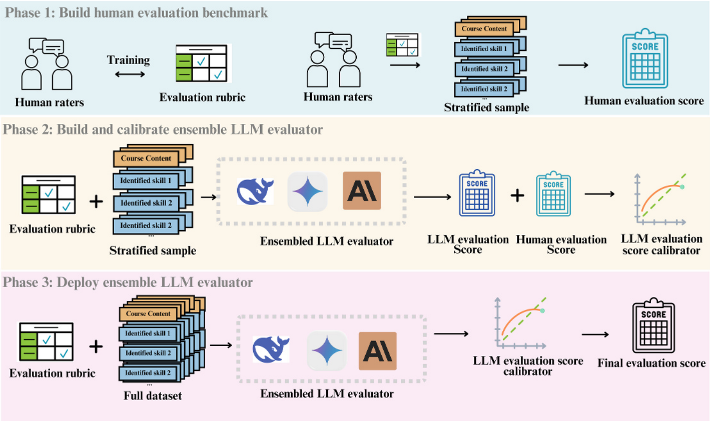

Fig. 1. Human-LLM collaborative evaluation framework

Human-LLM Collaborative Evaluation Framework. To assess the actual alignment of each extracted skill in a scalable manner, we build a human-LLM collaborative evaluation framework that combines human expertise and the power of LLMs, including three phases (Fig. 1):

1. **Build human evaluation benchmark.** Three annotators first score skills extracted from 10 courses across all methods to iteratively refine and finalize a scoring rubric (Table 1). Two annotators then apply the rubric to skills from 60 randomly selected curricula, resolving discrepancies through discussion and re-scoring until acceptable reliability was achieved (Cohens Kappa =

[2] The number 20 aligns with the typical number of DWAs associated with each occupation in *O*NET*.

0.747, ICC = 0.862). Finally, the two annotators score skills from another 60 stratified samples across data types, subjects, and alignment strategies, again resolving any differences to produce the finalized human evaluation benchmark dataset.

2. **Build and calibrate ensemble LLM evaluator.** Drawing on LLM-as-a-judge frameworks in the NLP domain [22], we use several state-of-the-art reasoning models DeepSeek R1, Gemini 1.5 Pro, and Claude 3.5 Sonnet as an ensemble evaluator. Each model scores the skills using few-shot prompts based on our rubric, and their scores are averaged. To improve reliability, we train a calibration model that predicts human scores from LLM outputs using linear regression with quantile interpolation. This helps correct bias and aligns the distribution of LLM predictions with human evaluations. The human evaluation benchmark dataset is split 80% and 20% for training and testing, using 10-fold cross-validation. The calibrated model achieves an accuracy of 0.709, a weighted Cohen's Kappa of 0.767, and a Krippendorff's alpha of 0.761, indicating good consistency with human judgments.
3. **Deploy ensemble evaluator.** Lastly, we deploy the calibrated LLM evaluator to assess the top 10 skills identified by each of the 10 alignment methods across 400 curriculum samples. Each extracted skill is independently scored using few-shot prompting, and the ensemble outputs are then calibrated using the model trained in Phase 2.

Table 1. Rubric for evaluating the actual alignment of each extracted skill

Score	Criteria
5	Core learning objective of the course; explicitly covered.
4	Aligns with the course; students should be able to perform it after completion.
3	Not explicitly covered, but transferable skills may be developed.
2	Within the same domain, but not directly relevant.
1	Outside the scope of the course; belongs to a different domain.

Performance Metrics. The skill alignment scores generated above are further aggregated into four metrics to evaluate the overall performance of each text alignment strategy: (1) $Precision_5$: % of top 10 extracted skills that score 5; (2) $Precision_4$: % of top 10 extracted skills that score 4 or higher; (3) $Mean$: Average alignment score of the top 10 skills; (4) Normalized Discounted Cumulative Gain($NDCG$): A metric that evaluates the ranking accuracy of information retrieval systems, with scores ranging from 0 to 1. Higher values indicate more accurate rankings [2].

3 Results

Table 2 summarizes the overall performance of each alignment method across the full dataset. RAG consistently outperforms both zero-shot and traditional methods in terms of extraction precision. Among zero-shot prompting methods, only the best-performing model, GPT-4o, surpasses traditional NLP approaches, while the average performance of zero-shot methods remains lower than traditional methods.

Table 2. Overall performance across the entire dataset

	TF-IDF	BERT	ZERO-SHOT				RAG			
			GPT	Llama	Claude	Gemini	GPT	Llama	Claude	Gemini
$Precision_5$	0.100	0.043	0.244	0.032	0.032	0.055	0.540	0.432	0.416	0.432
$Precision_4$	0.269	0.240	0.418	0.116	0.160	0.199	0.820	0.715	0.695	0.721
$Mean$	2.418	2.344	2.981	1.824	2.074	2.216	4.268	3.985	3.946	3.993
$NDCG$	0.887	0.878	0.868	0.881	0.869	0.899	0.959	0.971	0.973	0.973

We further examine performance heterogeneity across different types of curriculum documents, focusing on $Precision_4$, as identifying relevant skills is typically prioritized in skill extraction research. As shown in Fig. 2, RAG consistently outperforms both traditional NLP and zero-shot methods across all curriculum document types. In general catalogs, RAG achieves 47.2-59.8% precision, correctly identifying about half of the top 10 skills. In contrast, traditional NLP methods and most zero-shot models (except GPT-4o) score below 10%, often missing all relevant skills. For syllabi, traditional methods like TF-IDF and BERT perform better than most zero-shot models, identifying around 2.8-3.8 relevant skills, while others find only about 2.4. RAG again performs best, with 7.6-9.1 relevant skills on average.

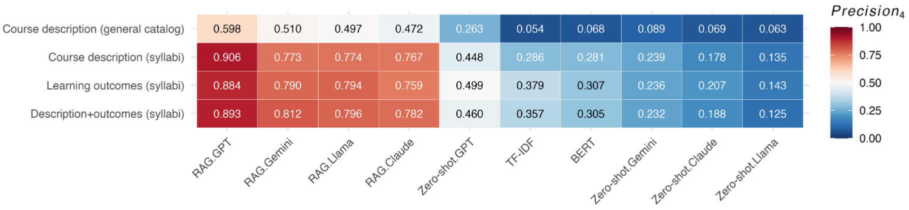

Fig. 2. $Precision_4$ comparison across different types of curriculum documents and LLM models

4 Discussion and Conclusion

In this study, we present one of the first systematic evaluations of text alignment strategies for the skill extraction task in CA, comparing traditional NLP methods and LLM-based approaches across different types of curriculum documents. Our findings reveal key insights into the overall performance and generalizability of these methods.

LLM-based methods, especially those using RAG, consistently outperform traditional approaches in alignment quality, precision, and ranking accuracy. This advantage holds across all types of curriculum documents. In particular, LLMs show strong improvements in handling brief, general documents, such as open catalogs, where traditional methods often struggle due to limited detail and the pedagogical reasoning required to address granularity mismatches. These findings underscore the promise of LLMs in addressing longstanding challenges in CA, especially when dealing with sparse or heterogeneous educational data.

We also examined how LLM performance varies across different models and document types. In zero-shot settings, performance differed notably between open-source and proprietary models, as well as by model size, parameters, and optimization goals. However, using RAG helped reduce this variation, resulting in more stable outcomes across models. Additionally, zero-shot prompting worked better on the most difficult curriculum document types for traditional methods, like general catalogs, but was less effective on structured, information-rich documents.

Our findings have several practical implications. First, LLMs can be a powerful alternative to traditional NLP methods when working with low-information, highly summarized curriculum documents. Second, while LLMs are promising for empowering CA tasks, effective use requires careful design and tuning beyond zero-shot prompting alone. Third, given the performance variation across prompting and model selection, careful evaluation, transparent reporting of methodological choices, and validation of LLM-involved analyses are crucial to ensuring the rigor and trustworthiness of research conclusions.

Supplemental Information

Supplemental Information can be accessed at: https://github.com/AEQUITAS-Lab/Evaluation-of-LLM-in-CA-AIED-2025

Acknowledgements. This work is supported by a Faculty Collaboration Grant from Teachers College, Columbia University. We also thank Joe Karaganis for providing access to Open Syllabus data.

References

1. Burrus, J., Jackson, T., Xi, N., Steinberg, J.: Identifying the most important 21st century workforce competencies: An analysis of the occupational information network (o* net). ETS Research Report Series **2013**(2), i–55 (2013)

2. Busa-Fekete, R., Szarvas, G., Elteto, T., Kégl, B.: An apple-to-apple comparison of learning-to-rank algorithms in terms of normalized discounted cumulative gain. In: ECAI 2012-20th European Conference on Artificial Intelligence: Preference Learning: Problems and Applications in AI Workshop, vol. 242. Ios Press (2012)
3. Chauhan, R.S.: Occupation exploration: Using O* NET in the management classroom. Manage. Teach. Rev. **4**(1), 79–88 (2019)
4. Chou, C.Y., et al.: Open student models of core competencies at the curriculum level: using learning analytics for student reflection. IEEE Trans. Emerg. Top. Comput. **5**(1), 32–44 (2015)
5. Deming, D.J.: The growing importance of social skills in the labor market. Q. J. Econ. **132**(4), 1593–1640 (2017)
6. Ehara, Y.: A support system to help teachers design course plans conforming to national curriculum guidelines. In: Artificial Intelligence in Education, vol. 1831, pp. 549–554. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-36336-8_85
7. Gottipati, S., Shankararaman, V.: Competency analytics tool: analyzing curriculum using course competencies. Educ. Inf. Technol. **23**, 41–60 (2018)
8. Hilliger, I., Miranda, C., Celis, S., Pérez-Sanagustín, M.: Curriculum analytics adoption in higher education: a multiple case study engaging stakeholders in different phases of design. Br. J. Edu. Technol. **55**(3), 785–801 (2024)
9. Hilton, M.L., Tippins, N.T.: A database for a changing economy: Review of the occupational information network (O*NET) (2010)
10. Javadian Sabet, A., Bana, S.H., Yu, R., Frank, M.R.: Course-skill atlas: a national longitudinal dataset of skills taught in us higher education curricula. Sci. Data **11**(1), 1086 (2024)
11. Jovanović, J., Zamecnik, A., Barthakur, A., Dawson, S.: Curriculum analytics: Exploring assessment objectives, types, and grades in a study program. Educ. Inform. Technol. 1–24 (2024)
12. Kawintiranon, K., Vateekul, P., Suchato, A., Punyabukkana, P.: Understanding knowledge areas in curriculum through text mining from course materials. In: 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), pp. 161–168 (2016). https://doi.org/10.1109/TALE.2016.7851788
13. Kwak, Y., Pardos, Z.A.: Bridging large language model disparities: skill tagging of multilingual educational content. Br. J. Edu. Technol. **55**(5), 2039–2057 (2024). https://doi.org/10.1111/bjet.13465
14. Li, X., Henriksson, A., Duneld, M., Nouri, J., Wu, Y.: Supporting teaching-to-the-curriculum by linking diagnostic tests to curriculum goals. In: Artificial Intelligence in Education. vol. 14829, pp. 118–132. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-64302-6_9
15. Light, J.: Student demand and the supply of college courses. Available at SSRN 4856488 (2024)
16. Malik, R., Abdi, D., Wang, R., Demszky, D.: Scaling high-leverage curriculum scaffolding in middle-school mathematics. In: Proceedings of the Eleventh ACM Conference on Learning @ Scale, pp. 476–480 (2024). https://doi.org/10.1145/3657604.3664698
17. Nguyen, K.C., Zhang, M., Montariol, S., Bosselut, A.: Rethinking skill extraction in the job market domain using large language models. In: 1st Workshop on Natural Language Processing for Human Resources. NLP4HR 2024, pp. 27–42. Association for Computational Linguistics, ACL Anthology (2024)

18. Noveski, G., Jeroncic, M., Velard, T., Kocuvan, P., Gams, M.: Comparison of large language models in generating machine learning curricula in high schools. Electronics **13**(20), 4109 (2024). https://doi.org/10.3390/electronics13204109
19. Pistilli, M.D., Heileman, G.L.: Guiding early and often: Using curricular and learning analytics to shape teaching, learning, and student success in gateway courses. N. Dir. High. Educ. **2017**(180), 21–30 (2017)
20. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992 (2019)
21. Tan, C.W., Lim, K.Y.: Revolutionizing formative assessment in STEM fields: Leveraging AI and NLP techniques. In: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1357–1364 (2023). https://doi.org/10.1109/APSIPAASC58517.2023.10317226
22. Wang, X., et al.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
23. Woessmann, L.: Skills and earnings: a multidimensional perspective on human capital. Ann. Rev. Econom. **17** (2024)
24. Yu, R., Das, S., Gurajada, S., Varshney, K., Raghavan, H., Lastra-Anadon, C.: A research framework for understanding education-occupation alignment with NLP techniques. In: Proceedings of the 1st Workshop on NLP for Positive Impact, pp. 100–106. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.nlp4posimpact-1.11
25. Zamecnik, A., Barthakur, A., Wang, H., Dawson, S.: Mapping employable skills in higher education curriculum using LLMs. In: European Conference on Technology Enhanced Learning, pp. 18–32. Springer (2024)