Unlocking Educational Resources with Colleague-IRGen: Personalized and Efficient Lesson Plan Search & Generation

Anonymous Author(s)

ABSTRACT

This paper presents *ColleagueIRGen*, an Artificial Intelligence (AI)-based system that aims to improve the efficiency and quality of lesson plan development for K-12 education. The system combines information retrieval (IR) and generative AI models with domain-specific educational knowledge to overcome the limitations of manual and general-purpose AI tools in creating pedagogically sound and personalized lesson materials. The system has demonstrated superior capability in accurately retrieving and generating relevant educational content through experiments and user studies. The results highlight the system's effectiveness in enhancing educational outcomes and emphasize the importance of domain-specific IR and AI integration in educational content creation. Also, this study identifies critical areas for future enhancements and paves the way for more personalized and inclusive educational experiences.

CCS CONCEPTS

• **Information systems** → **Information systems applications**; **Users and interactive retrieval**; • **Applied computing** → **Education**; • **Human-centered computing** → *Collaborative and social computing design and evaluation methods*; • **Social and professional topics** → **K-12 education**.

KEYWORDS

IR in Education, Domain-specific application, Semantic search

ACM Reference Format:

Anonymous Author(s). 2024. Unlocking Educational Resources with Colleague-IRGen: Personalized and Efficient Lesson Plan Search & Generation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (SIGIR'24).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The adoption of AI technologies such as ChatGPT [37], Bard [14], and Microsoft Copilot [34] by K-12 teachers is reshaping instructional tasks including material creation and grading, despite challenges in aligning AI-generated content with specific educational needs and pedagogical depth [9]. Despite the proliferation of online learning resources and technology in classrooms, lesson planning remains predominantly a manual and labor-intensive task, compounded by the demands of integrating diverse content and addressing the varied needs of students [9]. This essential process is crucial for effective teaching and engaging students. Still, it is hindered by the general-purpose AI tools currently in use for designing educational materials, by their inability to meet specific academic needs and conceptual and pedagogical depth, leading to inaccuracies and suboptimal instructions [9]. This complexity emphasizes the necessity for specialized systems that complement rather than replace human creativity, agency, and expertise [9]. *Colleague-IRGen* is introduced as an AI-enhanced system that facilitates the development of inclusive, high-quality lesson materials for K-12 educators, utilizing advancements in IR and generative AI.

The significance of lesson planning cannot be overstated, as it shapes student engagement and helps navigate the complexities of classroom dynamics. The excessive time and cognitive load involved in lesson preparation are particularly taxing for early-career teachers and those in challenging environments [23–26, 44, 48]. The diverse needs of students, including cultural, linguistic, and special educational requirements, further complicate this task of creating high-quality instructional materials and increase the cognitive load [16, 21, 27, 40]. Although Open Education Resources (OER) materials under Creative Commons License [e.g. 8, 20, 53] unlock access to instructional materials, the quality of these materials is not always guaranteed [16]. *Colleague-IRGen* stands out by leveraging IR, generative AI, and education domain-specific knowledge to enhance the efficiency and quality of lesson planning.

1.1 Background and Related Works

While existing research has concentrated on curriculum materials and textbooks [45, 47], we emphasize the significance of high-quality lesson plans in enhancing instruction and student achievement. Integrating generative IR and AI into lesson planning presents challenges, notably the limitations of general AI techniques and pre-trained Large Language Models (LLMs) to meet the nuanced needs of educational content. While applications like ChatGPT [37] provide broad functionalities, they often lack the subject-specific focus required for effective educational content creation [55]. On the other hand, traditional education resource retrieval and sharing platforms [e.g. 1, 8, 15, 20, 50, 52, 53], while rich in content, do not sufficiently support personalized lesson planning or ensure the embedding of pedagogical principles and content quality [55]. Similarly, newer AI-based tools for lesson planning [e.g. 30, 33, 51] demonstrate innovative features. Still, their alignment with research evidence and learning standards is yet to be fully explored.

The uniqueness of the educational domain necessitates specialized IR systems. These systems must account for educators' and learners' diverse, complex, and specific information needs. Recent educational content retrieval and recommendation advancements highlight the potential to transform lesson planning tasks. Advancements in domain-specific and context-aware IR systems [2, 3, 19, 43] show promise in transforming lesson planning by tailoring search

SIGIR'24, June 03–05, 2018, Woodstock, NY

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX
https://doi.org/XXXXXXX.XXXXXXX

results to specific educator needs and automating lesson plan generation. The potential of LLMs in predicting user preferences [54], retrieval-augmented generation (RAG) [e.g. 5, 13, 17, 31] and generative IR [e.g. 7, 32, 35] further highlight the utility of IR and AI approaches in providing contextually relevant lesson plans when fine-tuned with domain-specific knowledge [56].

Therefore, our research synthesizes domain-specific knowledge with IR and generative AI approaches to adapt technology for educational contexts, drawing parallels with advancements in various professional fields [28, 56]. *Colleague-IRGen* is designed to overcome the limitations of both AI-based educational resource generation applications [e.g. 30, 33, 51] and traditional resource retrieval and sharing platforms [e.g. 1, 8, 15, 20, 50, 52, 53] through a unique integration of IR and generative AI with educational domain expertise. This integration not only enhances lesson planning quality and efficiency but also incorporates rigorous vetting processes to ensure the reliability of lesson plans [9]. Furthermore, our methodology emphasizes human-centered AI development and principles of learning science. Addressing the scarcity of high-quality, large-scale training data, we have established a framework for lesson material quality, including 40 measures of rigor, engagement, and inclusivity, and annotated over 1000 lesson materials. This initiative enriches the dataset for algorithm development and evaluation, pushing forward the application of AI in education. Our goal is to push the application of AI in education to benefit educators and students.

2 COLLEAGUE-IRGEN WORKFLOW

Lesson planning is fundamental to effective classroom instruction, establishing the basis for comprehensive student problem-solving experiences. It involves the careful selection and sequencing of activities that adhere to core learning standards (e.g., [11]), encompass discrete knowledge and skills, and accommodate the diverse needs and learning styles of students [10]. *Colleague-IRGen* represents an AI-enhanced system designed to streamline lesson planning for K-12 educators by integrating generative IR with partial RAG specifically for educational content creation. Central to its functionality is the semantic search capability that utilizes contextual embedding to match educators' queries with a comprehensive database of lesson plans, ensuring selections meet high standards for academic rigor and inclusivity [10]. When an exact match is unavailable, the system's RAG component generates or refines lesson plans to closely align with the teacher's instructional needs, blending the benefits of IR and generative AI to simplify lesson preparation. Figure 1 presents an overview of the integrated workflow.

The workflow begins with user query submission, offering options for natural language or advanced structured inputs, facilitating both broad and granular search capabilities. Upon submission, the system interprets the query's intent through semantic analysis, transforming it into a vectorized form for comparison against stored lesson plans based on cosine similarity. This process ensures retrieval of highly relevant materials or triggers the lesson plan modification module for refinement or new generation when matches fall below a set similarity threshold.

Our lesson plan retrieval module leverages semantic search and advanced embedding techniques for understanding query intent,

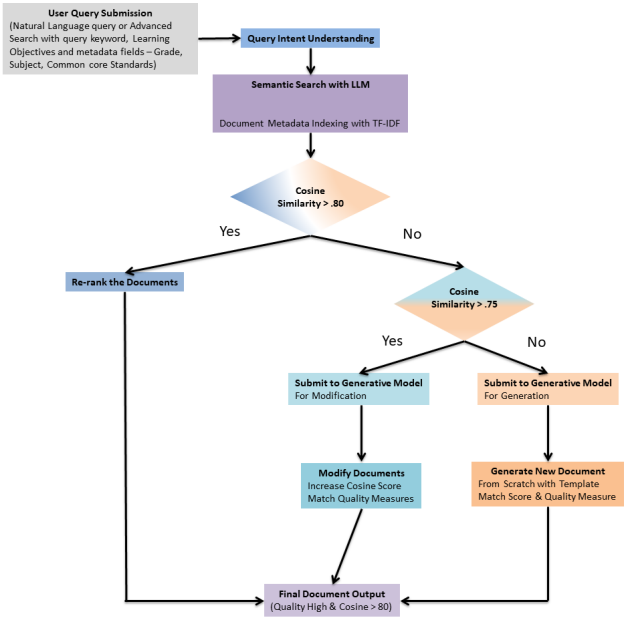

Figure 1: IR and generative AI workflow with decision points

effectively processing natural language and structured query components such as concepts, grades, subjects, and educational standards into vectorized forms. The module discerns the educational intent of queries by utilizing pre-trained embedding models (e.g., OpenAI's *text-embedding-ada-002* [36]) for lesson plan content and TF-IDF (Term Frequency-Inverse Document Frequency) [42] for structured query components. It then determines relevance by calculating cosine similarity scores between the query and lesson plan vectors in our database, prioritizing plans exceeding a certain similarity score threshold (e.g., .85).

The lesson plan modification module is designed to refine lesson plans to a higher degree of relevance when they fall below an optimal similarity threshold while retrieving lesson plans. Triggered for lesson plans whose cosine similarity score is below a certain threshold, this module selects the plan with the highest score for precision refinement rather than complete re-creation. The refinement phase is initiated by segmenting the lesson plan into specific sections using a pattern-matching algorithm, enabling precise modifications through prompt engineering with zero-shot learning in advanced LLMs (e.g., GPT-3.5-turbo [38] or GPT-4 [39]). This approach allows for focused improvements tailored to the lesson's needs. This refined process aims to realign the lesson plan with the educator's query and the educational standards by iteratively improving the content while maintaining its foundational structure and pedagogical integrity. The result is a lesson plan that matches the educator's criteria and is ready for classroom use, preserving the educational value and engagement of the original content.

Colleague-IRGen activates its lesson plan generation process when user queries do not closely match existing plans, indicated by a low cosine score (e.g., below .75). Instead of modifying available materials, it employs generative AI models (e.g., GPT-3.5-turbo [38] or GPT-4 [39]) and research-based templates to construct new lesson plans that adhere to educational standards and pedagogical

strategies, ensuring alignment with American K-12 education system's requirements, including state learning objectives, a sequence of activities for scaffolding learning and diverse student needs [10]. The detailed directive prompting technique is used for clarity and alignment with key lesson components: Warm Up, Explain, Reinforce, and Cool Down stages, creating a structured, coherent learning progress. They are organized in a sequence that reflects the logical flow of instruction, from introducing and explaining key concepts to reinforcing and summarizing them. This sequence mirrors the general structure of textbooks and pedagogical research. The following is an example prompt to generate a lesson plan for our evaluation experiments.

To maintain high lesson plan quality, Colleague-IRGen integrates a robust quality measurement framework developed through systematic literature review [12, 18, 41] and expert consultation, resulting in 40 measurable quality features across four domains - Rigor of Content, Engagingness of Activities, Inclusivity, and Differentiation, and Writing Quality [6]. In summary, the rigor of content quality is evaluated based on several criteria - alignment with learning standards, cognitive demand, mathematical terminology density, and accuracy. The effectiveness of activities is assessed based on the structure of the lessons, depth of questioning, discussions' orchestration, real-life context incorporation, multimedia use, and the presence of assessments. Inclusivity and differentiation are also considered to ensure that all students, regardless of disabilities, mastery levels, or language backgrounds, can participate. To improve clarity, writing quality is checked for essential meta-information, vocabulary diversity, and grammatical correctness. These quality measures are then simplified into a 5-star rating system for easy assessment. This framework guides the content generation process and ensures that generated lesson plans meet educational standards through iterative revisions.

The system's comprehensive workflow navigates from the semantic search for suitable lesson plans to employing generative models for new plans tailored to specific educational standards and user needs. This ensures the delivery of high-quality, pedagogically sound lesson plans customized to the educator's requirements, showcasing Colleague-IRGen's capability to enhance the lesson planning process through AI integration.

3 EVALUATION AND DISCUSSION

The evaluation of Colleague-IRGen encompassed a thorough two-phase process, focusing on its IR capabilities and, subsequently, on the quality and appropriateness of the generated lesson plans through a user study.

3.1 Creating Test Collection and Experiment

Our evaluation harnessed a comprehensive dataset of over 20,000 K-12 Mathematics lesson plans, sourced from various OER platforms such as IllustrativeMathematics [20] and BetterLesson [8], aligned with Common Core standards [11]. These plans, rich in narratives and multimedia elements, were standardized to include titles, learning objectives, and materials and structured into four key learning stages: Warm Up, Explain, Reinforce, and Cool Down. Each plan was meticulously aligned with Common Core State Standards for Mathematics (CCSSM) [11]. To assess the quality of these lesson plans, we enlisted four expert coders to evaluate a randomly selected subset of 1,000 lessons on 40 distinct quality features mentioned above. After a comprehensive four-week coding regimen that ensured accuracy and consistency, we achieved significant inter-rater reliability (Krippendorff's Alpha between 0.73 and 1), affirming the robustness of our evaluation methodology. We also intend to open-source the annotated lesson plans to the research community.

Furthermore, to simulate realistic teacher search scenarios in the absence of actual teacher queries, we generated synthetic queries drawing from our database of annotated lesson plans. The selection of a 10% sample for this purpose was strategically based on a random but stratified approach, ensuring a representative mix of grades, subjects, and Common Core Standards to closely mimic the diversity of real-life educational queries. This method, informed by protocols from [4, 22], leveraged GPT-3.5-turbo [38] to generate five distinct synthetic queries per lesson plan. These queries aimed to encapsulate a wide range of teacher intentions and instructional focuses. Upon generating these queries, we assessed their relevance to the lesson plans using cosine similarity measures, creating a test set that offers a comprehensive evaluation environment for our system's retrieval and generation capabilities.

Another vital aspect of simulating our experiment is worth mentioning in the selection of a cosine similarity threshold for our workflow decision-making points to activate IR, modification, and generation module and for determining lesson plan relevance, which was rooted in optimizing the balance between precision and recall in our system. These thresholds were established after multiple iterative processes, where we analyzed the trade-offs between retrieving highly relevant lesson plans and minimizing the inclusion of less pertinent results. For example, setting the threshold for retrieval at .85 allowed us to ensure that only lesson plans with high textual and conceptual similarity to the user's query were considered relevant.

3.2 Evaluation 1: IR System Performance and Ablation Study

Our first experiment evaluated Colleague-IRGen's retrieval effectiveness for Mathematics lesson plans, a challenging task due to the absence of direct educational IR benchmarks. Compared to traditional IR models such as TF-IDF [42], BM25 [46, 49], and COLBERT [29], Colleague-IRGen outperformed these models, achieving higher precision, recall, F1-Score, and accuracy. This superior performance can be attributed to specific aspects of Colleague-IRGen, such as its advanced semantic understanding and integration of educational domain knowledge, which enable more accurate identification and retrieval of content aligned with Common Core Standards.

The ablation study delved into the impact of different GPT versions on the system's output. Comparing GPT-3.5-turbo and GPT-4 revealed that GPT-4's contextual understanding and natural language processing advancements significantly contributed to the observed improvements. GPT-4's superior ability to grasp complex educational contexts and generate more coherent, relevant lesson plans led to higher precision, recall, and F1-Score, as well as enhanced alignment with educational standards. These results

Table 1: Evaluation results for different IR models and Colleague-IRGen

IR Model	Precision	Recall	F1-Score	Accuracy
BM25	0.82	0.83	0.84	0.84
TF-IDF	0.82	0.89	0.85	0.87
ColBERT	0.87	0.90	0.92	0.92
Colleague-IRGen (GPT-3.5-turbo)	0.92	0.95	0.94	0.95
Colleague-IRGen (GPT-4)	0.95	0.97	0.96	0.97

highlight the critical role of evolving generative AI models in improving educational content creation, offering insights into future enhancements for Colleague-IRGen to better meet educators' needs and optimize user experience. Table 1 presents the performance of our experiment.

3.3 Experiment 2: User Study Insights

We conducted a small user study to assess the performance of our system's modification and generative lesson plan modules by evaluating 10 modified and 10 generated lesson plans by two domain experts. The experts assessed each plan based on relevance to the query, usefulness, organization and completeness, and curriculum standards alignment. Relevance to the query evaluated how well the lesson plans matched the educators' search intentions, ensuring content directly addressed specified needs. Usefulness measured the plans' practical applicability in classroom settings, indicating their potential to enhance teaching and learning. Organization and completeness examined the structure of the lesson plans and whether they included all components necessary for effective instruction. The alignment of Curriculum standards ensured the plans met required content and skill development benchmarks. One expert gave high ratings across these metrics, highlighting the system's efficacy in facilitating lesson planning. Meanwhile, another expert's slightly lower scores on relevance and curriculum alignment suggest areas for further enhancement. These results affirm the system's usability and compliance with educational standards, emphasizing the need for continuous improvements to meet the diverse needs of users. The study's outcomes are summarized in Table 2, which presents the average scores assigned by each evaluator across all 20 lesson plans (10 modified and 10 generated) for each metric, using a 5-point scale. This approach provides a consolidated view of the system's performance in relevant areas as judged by domain experts.

Table 2: User evaluation of the modified and generated lesson plans

Metric	Annotator A	Annotator B
Relevance	4.3	3.8
Usefulness	4.3	4.0
Organization and completeness	4.2	4.2
Alignment with Standards	4.0	4.0

The evaluation of *Colleague-IRGen* through a two-phase study demonstrates its capability to significantly enhance the efficiency of lesson planning. However, the experiments also highlight the need for more personalized content to meet varied educational goals and student abilities, emphasizing the necessity of future system enhancements. Challenges in ensuring content accuracy, pedagogical appropriateness, and curriculum standards compliance remain. These challenges highlight the limitations of general pre-trained models in education-specific applications without fine-tuning, as evidenced by the user study where the generative model occasionally failed to produce lessons aligned with Common Core Standards. This indicates the critical need for integrating educational domain knowledge more deeply into AI models and for ongoing system refinement informed by user feedback.

4 CONCLUSION AND FUTURE WORK

In conclusion, our study introduces *Colleague-IRGen*, a system designed to automate the retrieval and generation of lesson plans. By leveraging advanced IR, data generation techniques, and a united sequencing method, *Colleague-IRGen* effectively produces coherent and pedagogically sound lesson materials. Our evaluation, conducted using a comprehensive Mathematics corpus, demonstrates the system's proficiency in improving retrieval accuracy and user experience, thus significantly enhancing the efficiency of lesson planning.

However, our study is not without limitations. The reliance on a small user study sample, the potential for expert rating bias, and the experiments' focus on Mathematics highlight areas for future improvement. We plan to address these limitations through extensive future work, including a long randomized control trial to comprehensively assess *Colleague-IRGen*'s educational impact. Efforts will also focus on enhancing the system's capacity for personalizing lesson plans to accommodate varying student learning objectives and skill levels, alongside integrating more detailed user feedback to refine our approaches to lesson plan mining, retrieval, and generation. *Colleague-IRGen* currently retrieves or generates a single lesson plan for a given query. Therefore, a pivotal goal is to extend *Colleague-IRGen*'s capabilities to create lesson plans of varying difficulties per query, ensuring a broader adaptation to diverse educational standards and learning environments. Moving forward, we aim to broaden our research to encompass a broader user base and a more varied range of subject matter, striving to improve the system's adaptability and effectiveness across different educational contexts.

REFERENCES

[1] achievethecore.org. 2024. Achieve the Core. achievethecore. https://achievethecore.org/ Accessed: 2024-02-01.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Krishnaram Kenthapadi, Nitish Srivastava, and Raja Velu. 2010. Enriching textbooks through data mining. In *Proceedings of the First ACM Symposium on Computing for Development*. 1–9.

[3] Rakesh Agrawal, Behzad Golshan, and Evangelos Papalexakis. 2016. Toward Data-Driven Design of Educational Courses: A Feasibility Study. *Journal of Educational Data Mining* 8, 1 (2016), 1–21.

[4] Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can generative llms create query variants for test collections? an exploratory study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1869–1873.

[5] Avinash Anand, Arnav Goel, Medha Hira, Snehal Buldeo, Jatin Kumar, Astha Verma, Rushali Gupta, and Rajiv Ratn Shah. 2023. SciPhyRAG-Retrieval Augmentation to Improve LLMs on Physics Q &A. In *International Conference on Big Data Analytics*. Springer, 50–63.

[6] Anonymous. 2023. Anonymous. In *Anonymous*.

[7] Garbiel Bénédict, Ruqing Zhang, and Donald Metzler. 2023. Gen-ir@ sigir 2023: The first workshop on generative information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3460–3463.

[8] BetterLesson. 2024. BetterLesson. BetterLesson. https://betterlesson.com/about-us/ Accessed: 2024-02-01.

[9] Miguel A Cardona, Roberto J Rodríguez, Kristina Ishmael, et al. 2023. Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations. (2023).

[10] Jeffrey Choppin, Amy Roth McDuffie, Corey Drake, and Jon Davis. 2022. The role of instructional materials in the relationship between the official curriculum and the enacted curriculum. *Mathematical thinking and learning* 24, 2 (2022), 123–148.

[11] Council of Chief State School Officers. 2024. Common Core State State Standard Initiative. OpenAI. https://www.thecorestandards.org/ Accessed: 2024-02-01.

[12] Allison Davis, Robin Griffith, and Michelle Bauml. 2019. How preservice teachers use learner knowledge for planning and in-the-moment teaching decisions during guided reading. *Journal of Early Childhood Teacher Education* 40, 2 (2019), 138–158.

[13] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

[14] Google. 2024. Bard: A conversational AI tool by Google. Google. https://bard.google.com Accessed: 2024-02-01.

[15] Gooru Learning. 2024. gooru Navigator. gooru. https://goorulearning.com/about/ Accessed: 2024-02-01.

[16] Joshua T Hertel and Nicole M Wessman-Enzinger. 2017. Examining Pinterest as a curriculum resource for negative integers: An initial investigation. *Education Sciences* 7, 2 (2017), 45.

[17] Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. 2023. ChaTA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs. *arXiv preprint arXiv:2311.02775* (2023).

[18] Heather C Hill and Mark Chin. 2018. Connections between teachers' knowledge of students, instruction, and achievement outcomes. *American Educational Research Journal* 55, 5 (2018), 1076–1112.

[19] Chiao-Ching Huang, Tsung-En Huang, Chia-Hung Shih, Von-Wun Soo, and Yao-Chen Lin. 2010. Recommending a personalized lesson plan based on constraint satisfaction and negotiation. In *2010 2nd International Conference on Education Technology and Computer*, Vol. 1. IEEE, V1–12.

[20] Illustrative Mathematics. 2024. Illustrative Mathematics. Illustrative Mathematics. https://illustrativemathematics.org/ Accessed: 2024-02-01.

[21] Véronique Irwin, Ke Wang, Tabitha Tezil, Jijun Zhang, Alison Filbey, Julie Jung, Farrah Bullock Mann, Rita Dilig, and Stephanie Parker. 2023. Report on the Condition of Education 2023. NCES 2023-144. *National Center for Education Statistics* (2023).

[22] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query Expansion by Prompting Large Language Models. *arXiv preprint arXiv:2305.03653* (2023).

[23] John Jerrim and Sam Sims. 2019. The teaching and learning international survey (TALIS).

[24] John Jerrim and Sam Sims. 2021. When is high workload bad for teacher wellbeing? Accounting for the non-linear contribution of specific teaching tasks. *Teaching and Teacher Education* 105 (2021), 103395.

[25] Peter D John. 2006. Lesson planning and the student teacher: re-thinking the dominant model. *Journal of Curriculum Studies* 38, 4 (2006), 483–498.

[26] Nathan D Jones, Eric M Camburn, Benjamin Kelcey, and Esther Quintero. 2022. Teachers' time use and affect before and after COVID-19 school closures. *Aera Open* 8 (2022), 23328584211068068.

[27] Julia H Kaufman, V Darleen Opfer, Michelle Bongard, and Joseph D Pane. 2018. *Changes in what teachers know and do in the Common Core era.* Santa Monica, CA: RAND. https://www. rand. org/pubs/research_reports/RR2658 ....

[28] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024* (2022).

[29] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 39–48.

[30] LessonPlans.ai. 2024. LessonPlans.ai. LessonPlans.ai. https://www.lessonplans.ai/ Accessed: 2024-02-01.

[31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[32] Xiaoxi Li, Yujia Zhou, and Zhicheng Dou. 2023. UniGen: A Unified Generative Framework for Retrieval and Question Answering with Large Language Models. *arXiv preprint arXiv:2312.11036* (2023).

[33] magicschool.ai. 2024. magicschool.ai. magicschool.ai. https://www.magicschool.ai/ Accessed: 2024-02-01.

[34] Microsoft. 2024. Microsoft Copilot. Microsoft. https://www.microsoft.com/en-us/microsoft-copilot Accessed: 2024-02-01.

[35] Marc Najork. 2023. Generative information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–1.

[36] OpenAI. 2023. text-embedding-ada-002. https://openai.com/blog/new-and-improved-embedding-model. Accessed: 2024-02-02.

[37] OpenAI. 2024. ChatGPT: A conversational agent. OpenAI. https://openai.com/chatgpt Accessed: 2024-02-01.

[38] OpenAI. 202X. Introducing GPT-3.5-turbo. https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed: 2024-02-02.

[39] OpenAI. 202X. Introducing GPT-4. https://openai.com/research/gpt-4. Accessed: 2024-02-02.

[40] Morgan Polikoff. 2021. *Beyond standards: The fragmentation of education governance and the promise of curriculum reform.* Harvard Education Press.

[41] Morgan S Polikoff, Hovanes Gasparian, Shira Korn, Martin Gamboa, Andrew C Porter, Toni Smith, and Michael S Garet. 2020. Flexibly using the surveys of enacted curriculum to study alignment. *Educational Measurement: Issues and Practice* 39, 2 (2020), 38–47.

[42] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Citeseer, 29–48.

[43] Rishabh Ranawat, Ashwin Venkataraman, and Lakshminarayanan Subramanian. 2021. Collectiveteach: a system to generate and sequence web-annotated lesson plans. In *ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–13.

[44] Philip M Reeves, Wik Hung Pun, and Kyung Sun Chung. 2017. Influence of teacher collaboration on job satisfaction and student achievement. *Teaching and Teacher Education* 67 (2017), 227–236.

[45] Janine T Remillard. 1999. Curriculum materials in mathematics education reform: A framework for examining teachers' curriculum development. *Curriculum Inquiry* 29, 3 (1999), 315–342.

[46] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer, 232–241.

[47] William Schmidt, Richard Houang, and Leland Cogan. 2002. A coherent curriculum. *American Education* 26, 10 (2002), 1–18.

[48] Min Sun. 2018. Black teachers' retention and transfer patterns in North Carolina: How do patterns vary by teacher effectiveness, subject, and school conditions? *AERA Open* 4, 3 (2018), 2332858418784914.

[49] Krysta M Svore and Christopher JC Burges. 2009. A machine learning approach for improved BM25 retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 1811–1814.

[50] Teachers Pay Teachers. 2024. Teachers Pay Teachers. TPT. https://www.teacherspayteachers.com/ Accessed: 2024-02-01.

[51] teachology.ai. 2024. teachology.ai. teachology.ai. https://www.teachology.ai/ Accessed: 2024-02-01.

[52] The American Federation of Teachers. 2024. Share My Lesson. sharemylesson. https://sharemylesson.com/ Accessed: 2024-02-01.

[53] The Institute for the Study of Knowledge Management in Education. 2024. OER-Commons Open Education Resources. OERCommons. https://oercommons.org/about Accessed: 2024-02-01.

[54] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621* (2023).

[55] Dazhen Tong, Yang Tao, Kangkang Zhang, Xinxin Dong, Yangyang Hu, Sudong Pan, and Qiaoyi Liu. 2023. Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. *Asia Pacific Education Review* (2023), 1–11.

[56] Weiqi Xu and Fan Ouyang. 2022. The application of AI technologies in STEM education: a systematic review from 2011 to 2021. *International Journal of STEM Education* 9, 1 (2022), 1–20.