Agentic Workflows for Enhancing Course Development

Bartolf, D. E., Black, N., Bickett, K., Brown, V., Patton, B., & Schepper, E.

Online curriculum design often struggles to translate learning theory into effective instructional materials. This study applies design-based research methodology to develop and evaluate a novel curriculum builder framework powered by agentic workflows and Retrieval-Augmented Generation (RAG). The multi-agent architecture addresses key challenges by dynamically mapping prerequisite knowledge, generating contextually relevant content, and supporting Problem-Based Learning through specialized planning, writing, evaluation, and revision agents. Our framework systematically integrates pedagogical principles while minimizing teacher effort in content development. Qualitative and quantitative evaluation data from interviews, surveys, and platform analytics demonstrates significant improvements in student engagement, persistence, and confidence, alongside enhanced content quality and educator satisfaction. This approach provides a scalable, adaptable solution for creating learner-centered instructional materials that effectively bridge the gap between theoretical learning science principles and practical curriculum implementation.

Introduction

Despite the comprehensive theoretical frameworks offered by learning sciences for curriculum development, translating these theories into practical and effective instructional materials remains challenging (Ertmer & Newby, 2013; Domingo, 2024; Norman & Lotrecchiano, 2021). Key issues include clearly communicating prerequisite knowledge, effectively scaffolding learning pathways, and providing contextually relevant examples that resonate with diverse student populations.

To address these critical challenges, this paper examines three specific research questions:

1. How can AI-supported curriculum development frameworks bridge the gap between learning theory and practical curriculum design?
2. How does integrating AI-driven workflows, particularly agentic models and Retrieval-Augmented Generation (RAG), impact the quality, accuracy, and contextual relevance of educational content?
3. In what ways does the proposed AI curriculum builder framework affect teacher effort, student engagement, persistence, and learner confidence in online learning environments?

These questions are investigated by exploring the evolution and application of QuantHub's data literacy education platform, a leader in data science education that specializes in personalized asynchronous and synchronous e-learning.

QuantHub curriculum relied upon manual content development focused on incorporating foundational learning theories, including Cognitive Load Theory (CLT), which advocates structuring content to minimize cognitive overload (Sweller, 2011), and Elaboration Theory, which emphasizes a progression from simple to complex content (Reigeluth & Stein, 1983). The curriculum was delivered through adaptive learning pathways.

Evaluation of QuantHub's initial data literacy curriculum, delivered during a single academic year pilot period to high middle school and high school students, revealed practical limitations, particularly regarding novice learners' experiences. Challenges included inadequate personalized scaffolding, difficulty mapping clear learning pathways, and the absence of sufficiently relevant examples across various contexts to improve learner engagement (Jumaat & Tasir, 2014). To overcome these shortcomings, QuantHub sought to develop a novel AI-driven curriculum framework that integrates multi-agent workflows and Retrieval-Augmented Generation (RAG). The final curriculum would support authentic Problem-Based Learning (PBL) scenarios.

Background

Context and Challenges

Formal data from QuantHub's initial pilot—engaging students in STEM and non-STEM course—showed clear evidence of student disengagement. While 54% of students completed introductory modules like data protection concepts, completion rates sharply declined to below 6% for more complex analytics content (Snyder, 2023). Informal teacher observations provided additional qualitative insights, noting particular struggles among English language learners and students reading below grade level.

Platform analytics demonstrated significantly lower engagement when students perceived instructional materials as irrelevant to their daily lives or career aspirations. Initial modules intended for broad audiences had low interaction rates. Conversely, when curriculum developers integrated examples from popular culture, sports, and specific career pathways, platform data revealed improved engagement metrics. As one educator informally observed, "Connecting it to industry at key moments clearly engages students" (Cochran, 2024).

Additionally, data indicated significant challenges with student persistence: fewer than 25% of students accounted for approximately 80% of platform interactions (Snyder, 2023). These analytics provided a clear quantitative basis for concerns about uneven participation. Similarly, formal survey results indicated low student confidence, with 51–61% of respondents reporting limited understanding of key data literacy concepts (Snyder, 2023).

Resource Allocation and Content Development Challenges

Addressing these identified challenges required significant resource investment. Initial informal observations and formal content quality reviews identified substantial inconsistencies in curriculum created by independently working subject matter experts (SMEs). These informal insights were corroborated by formal quality assurance processes conducted by instructional designers, who noted issues in terminology, instructional coherence, and overall quality.

Response and Product Development

In response to both informal and formal findings regarding instructional challenges and resource constraints, the curriculum team sought to transition from manual SME-driven processes to an AI-assisted approach.

The intent was to develop a system of integrated AI tools that scaffolds learning activities, generates contextually relevant content, and creates authentic problem-based scenarios.

Literature Review

The integration of AI into educational systems has opened new avenues for curriculum development and personalized learning. This literature review examines the theoretical foundations and empirical research that inform the development of AI-enabled systems for educational content creation, particularly those utilizing agentic models. By analyzing current research on instructional design principles, AI capabilities, and their applications in educational contexts, this review establishes a framework for understanding how these technologies can improve accuracy, quality, and contextual relevance in curriculum content creation, providing the foundational knowledge that guided QuantHub's development of AI-assisted curriculum development.

Instructional Design Theory and Problem-Based Learning

Instructional design frameworks such as PBL offer robust methods for increasing student engagement, promoting critical thinking, and supporting meaningful learning through real-world problem-solving tasks. Rooted in constructivism, PBL emphasizes active student participation and deeper understanding compared to traditional instructional methods (Rotthoff et al., 2015). PBL incorporates scaffolding, systematically guiding learners from basic knowledge toward more complex tasks (Reigeluth & Stein, 1983). This structured approach is essential in online learning environments, supporting incremental skill growth and cognitive load management, critical to maintaining learner engagement (Jumaat & Tasir, 2014; Sweller, 2011).

Feedback mechanisms play a crucial role in PBL by helping learners explicitly link new knowledge to existing skills, thus enhancing retention and application of concepts (Fries et al., 2021; Chen, 2008). Additionally, the Practicing Connections Hypothesis (PCH) emphasizes that learning is most effective when theoretical concepts directly connect to practical applications, thereby enhancing skill transfer and professional readiness (Fries et al., 2021).

Despite these advantages, traditional PBL implementations often encounter scalability and personalization challenges, particularly in online contexts (Meng et. Al., 2023). Effective scaffolding and feedback are resource-intensive, and generic PBL models can struggle to adapt to diverse learner backgrounds and abilities, thus underscoring the need for scalable, adaptable solutions.

Technical Foundations of AI in Educational Systems

AI-based solutions, specifically agentic models, have emerged as effective means to address the scalability and adaptability challenges inherent in traditional PBL implementations. Agentic models are sophisticated AI systems capable of autonomous reasoning and dynamic adaptation to complex instructional tasks (Singh et al., 2024). Their strength lies in synthesizing educational standards, learner profiles, and pedagogical objectives to generate tailored instructional content. However, Singh et al. (2024) and Gillani et al. (2023) highlight significant limitations. Singh et al. (2024) acknowledge that despite the adaptability of agentic systems, their practical application in diverse classroom environments remains challenging due to computational resource demands and limited transparency in decision-making processes. The "black-box" nature of these models often hinders educator acceptance and effective integration into standard teaching practices (Gillani et al., 2023).

Retrieval-Augmented Generation (RAG)

Current research shows that RAG can significantly enhance the reliability of AI-generated content by integrating external, verified knowledge sources. Unlike traditional generative models that may produce inaccurate or irrelevant outputs ("hallucinations"), RAG provides a systematic grounding in domain-specific knowledge bases, thus significantly increasing content accuracy and reliability (Woo et al., 2024; Kahl et al., 2024). By introducing a quality assurance layer combining human oversight with automated metrics like ROUGE and METEOR scores--measures of content similarity and quality--Woo et al. (2024) demonstrated a 39.7% improvement in accuracy using RAG over traditional LLM outputs, clearly establishing its effectiveness in precision-sensitive domains like education and medicine. Kahl et al. (2024) similarly illustrated how multimodal materials integrated through RAG markedly reduce inaccuracies in educational contexts.

Woo et al. (2024) and Kahl et al. (2024), although validating RAG's improvement over traditional generative models, point out limitations regarding the efficiency of RAG-based systems. These include latency issues and computational costs, especially when integrating multimodal sources or extensive external knowledge bases, and a reliance upon a well-defined human-system collaboration. Furthermore, Woo et al. (2024) emphasize the necessity of continuous updates to external databases, which can be resource-intensive and may lead to maintenance challenges.

Technical Synthesis and Challenges for Data Science Education

Building upon the RAG foundational architectures, Mitra et al. (2024) expanded the scope of RAG through an agentic framework that specifically addresses challenges in data science education. Their framework automates the creation, refinement, and validation of synthetic datasets, which is particularly valuable for teaching data analytics concepts. Data science curricula often require diverse, realistic datasets that illustrate specific statistical properties or domain characteristics. By leveraging synthetic data generation alongside traditional retrieval mechanisms, their framework enables students to encounter varied data scenarios while maintaining instructional control over the learning progression.

Fact-Checking Frameworks

Ensuring the validity and trustworthiness of content is essential, particularly for curriculum builders tasked with generating instructional materials. Recent advancements in agentic AI offer insights into designing robust frameworks that prioritize accuracy and transparency. Li et al. (2024) developed FactAgent, a structured AI workflow integrating internal linguistic analysis and external credibility checks, dramatically enhancing the reliability of generated content.

Domain-Specific Specialization in Educational AI

The use of agentic AI in academic curriculum development has significant potential to improve content validity by integrating domain-specific expertise and enabling collaboration across disciplines. Domain-specific fine-tuning and specialization, as illustrated by Aryal et al. (2024) and Chu et al. (2024), enhance the quality of educational AI outputs by integrating

interdisciplinary expertise and dynamic, context-sensitive adaptation capabilities. Aryal et al.'s multi-agent systems and Chu et al.'s Professional Agents (PAgents) provide robust methods for continually refining AI knowledge, ensuring its relevance and precision across diverse instructional contexts.

While research demonstrates that value of AI, Aryal et al. (2024) and Chu et al. (2024) stress that domain-specific fine-tuning, while improving relevance, may inadvertently limit AI adaptability across interdisciplinary or evolving educational contexts. Chu et al. (2024) specifically caution about over-specialization, noting that highly tailored systems risk diminished performance when encountering content outside their trained domains or when pedagogical paradigms shift.

Human-AI Collaboration Frameworks

Human-AI collaborative frameworks, as proposed by Dierickx et al. (2024), emphasize transparency, explainability, and iterative refinement to maintain trust and accuracy in AI-supported educational materials. Recent reviews further suggest that effective human-AI collaboration remains challenging, particularly regarding accurately capturing and integrating nuanced pedagogical expertise (Dierickx et al., 2024). Their findings highlight the necessity of more sophisticated interfaces and clearer mechanisms for educators to effectively oversee and influence AI-driven curriculum development processes.

Both Aryal et al. (2024) and Chu et al. (2024) studies emphasize the importance of collaboration among agents and the integration of human feedback. Aryal et al.'s system employs a structured orchestration mechanism where agents work in a predefined sequence, enabling efficient task execution. Chu et al. build upon this concept by introducing coevolution and human-in-the-loop feedback, allowing agents to learn from one another and refine their capabilities over time.

This reinforces the findings of Woo (2024), Singh et al. (2024), and Gillani et al. (2024) and highlights a critical consideration for educational applications: while AI can significantly enhance content creation efficiency, human expertise remains essential for ensuring pedagogical alignment and contextual appropriateness.

Application Synthesis and Implications for Curriculum Development

These advancements have important implications for AI-enabled curriculum development. By incorporating agentic models, RAG, knowledge graphs, and specialized multi-agent systems, curriculum developers may be able to create educational materials that are accurate, relevant, and adaptable to support problem-based learning design and delivery. The integration of external knowledge bases and structured validation workflows could ensure that content aligns with problem-based learning approaches. Furthermore, human-AI collaboration could enhance the reliability and trustworthiness of the generated content, fostering credibility in AI-enabled educational tools.

Methodology

This study employs design-based research (DBR) methodology to develop and evaluate an AI-supported curriculum builder. DBR is particularly appropriate for this work as it allows for the systematic design of educational innovations based on theoretical principles, iterative testing in authentic contexts, and refinement based on empirical findings (Wang & Hannafin, 2005). contribute to theoretical understanding of AI-enhanced instructional design processes.

The design-based research process consisted of three interconnected phases: (1) analysis of practical problems through literature review and needs assessment; (2) development of the AI curriculum builder informed by theoretical principles; and (3) iterative cycles of testing and refinement (based on Reeves, 2006).

The following sections detail each phase of this process, beginning with how findings from the analysis phase informed the technical design of the curriculum builder system, followed by the implementation approach and evaluation methods.

Analysis Phase: Problem Identification and Needs Assessment

The design-based research began with a comprehensive analysis of curriculum development challenges and needs assessment across multiple stakeholders. This phase integrated findings from the literature review with empirical data collected from QuantHub's existing platform implementation.

Key Challenges Identified from Literature Review

The literature review revealed several persistent challenges in curriculum development for data science education. Retrieval-Augmented Generation (RAG) research by Woo et al. (2024) and Kahl et al. (2024) highlighted the critical need for grounding AI-generated educational content in authoritative sources to prevent hallucinations and ensure factual accuracy. Singh et al. (2024) identified limitations in traditional linear LLM interactions, demonstrating the potential for multi-agent systems to improve consistency and performance in complex tasks like curriculum development. Additionally, studies on domain-specific agent specialization by Aryal et al. (2024) and Chu et al. (2024) emphasized the importance of expert-level knowledge in creating instructionally sound educational materials.

Needs Assessment Process

To complement the literature findings, the following systematic needs assessment uses multiple data sources from QuantHub's existing platform:

1. Analysis of Platform Usage Data: Completion rates across different modules and content types identified significant drop-offs in student progression. As documented previously, completion rates declined dramatically from 54% for introductory data protection concepts to less than 6% for advanced analytical content (Snyder, 2023).
2. Teacher Interviews and Observations: Qualitative feedback from educators implementing the platform, documenting their observations of student engagement patterns and challenges revealed specific issues with content relevance and cognitive accessibility, as exemplified by one teacher's observation that "The students are overwhelmed [and] will decide to just choose something and get it wrong, if need be, rather than read the content" (Cochran, 2024).
3. Content Development Process Analysis: The existing content creation workflow documentation revealed inefficiencies in the collaboration between system designers and subject matter experts, which resulted in costly delays and inconsistent quality.

Curriculum Design Needs Identified

The needs assessment revealed specific requirements across three key stakeholder groups:

1. Content Developers' Needs:
    1. Reduction in time-intensive manual content creation and revision cycles
    2. Improved consistency in terminology and instructional design across contributors
    3. More efficient coordination between system designers and subject matter experts
    4. Ability to create contextually relevant examples across diverse subject domains

1. Teachers' Needs:
    1. Materials that provide appropriate scaffolding for novice learners
    2. Content for self-paced and teacher-facilitated learning contexts

    3. Learning materials that maintain student engagement through relevant contexts

1. Learners' Needs:
    1. Scaffolded progression from simple to complex concepts
    2. Content presented in contexts relevant to their interests and future aspirations
    3. Supports for building confidence in applying data science concepts

Synthesis of Findings Informing Design Decisions

The analysis phase yielded several key insights that directly informed system design decisions:

1. Need for Dynamic Content Adaptation: The significant variation in completion rates and teacher feedback demonstrated that one-size-fits-all content was ineffective. This led to the decision to develop a multi-agent system capable of generating content tailored to specific learner groups.
2. Importance of Contextual Relevance: Student engagement improved when content incorporated examples from domains relevant to learners' interests. This informed the development of context-generation capabilities within the curriculum builder.
3. Critical Role of Scaffolding: The challenges faced by novice learners highlighted the need for systematic personalized scaffolding. This guided the implementation of evaluation agents that specifically check for appropriate progression from simple to complex concepts.
4. Resource Constraints in Manual Development: Inefficiencies in the content development process, including the challenges of maintaining consistency across multiple subject matter experts, pointed to the need for an AI-supported system that could maintain quality while reducing development time.
5. Verification and Accuracy Requirements: Given the technical nature of data science education, the system needed robust mechanisms to ensure factual accuracy. This finding directly influenced the decision to implement RAG with corrective mechanisms (CRAG) to ground content in verified information.

These findings collectively established the design requirements for the AI curriculum builder system, directly connecting the challenges identified in the literature review and needs assessment to the technical architecture developed in the next phase. By grounding the design decisions in both theoretical principles and empirical data from existing platform implementation, it ensured that the system would address real educational challenges while advancing the integration of AI in curriculum development.

Development Phase: Designing the AI Curriculum Builder

Based on the challenges identified in the analysis phase, AI curriculum builder was designed leveraging a multi-agent large language model (LLM) system incorporating planning, writing, and evaluation capabilities. This phase focused on addressing the limitations of traditional linear LLM interactions by employing a low-level, highly customizable framework to support the development of effective data science curriculum materials.

Low-Level Customizable Framework Development

The decision to use a low-level, highly customizable framework stemmed from the need for granular control over the AI interactions within the agentic model. Traditional low-code and no-code solutions often abstract away critical components of the AI system, resulting in a loss of control over essential parameters such as prompts, model selection, temperature settings, output formats, and memory management. By architecting the curriculum builder on a low-level framework, developers can fine-tune these parameters to optimize the performance and reliability of the AI-generated content.

This approach directly addresses the consistency challenges identified in the needs assessment, where the previous content development process resulted in variable quality across different subject matter experts. The framework implementation

aligns with Aryal et al.'s (2024) research on multi-agent systems, which demonstrated that tightly controlled orchestration mechanisms are essential for maintaining coherence in complex knowledge domains like data science education.

In the initial stages, the development team recognized that while large language models can generate high-quality outputs, they still face limitations when relying solely on parametric knowledge. The LLM was unable to generate accurate output based on discoveries after its training (required for courses that focused on teaching about emerging technologies), and it lacked the ability to tailor outputs to specific or highly contextualized situations, required for targeted case studies. This realization aligns with findings by Singh et al. (2024), who highlight the challenges LLMs face in maintaining consistency and accuracy without external grounding. To address this, the team designed a system that not only utilizes LLMs but also incorporates planning, writing, and evaluation capabilities within an agentic framework.

Multi-Agent System Design

The core of the curriculum builder is a multi-agent system where each agent specializes in a specific task, using an LLM as its "brain" and having access to specialized tools and resources. This system mirrors a team or squad working towards a common goal, with each member contributing their expertise to achieve optimal results.

The multi-agent architecture directly implements the domain specialization approach advocated by Chu et al. (2024), who introduced the concept of Professional Agents (PAgents) with specialized modules for perception, reasoning, and action. Each of the system agents embodies this modular approach, focusing on specific aspects of curriculum development. The separation of planning, writing, and evaluation functions specifically addresses the finding from the needs assessment that the content review process created significant bottlenecks in the previous manual workflow.

Agents and Their Specializations. The agents in the system include:

Interview agent. This agent interfaces with the user, gathering initial information through dynamic and unscripted conversations. It acts as a consultant, asking questions that are not hard-coded or predetermined, and continues the dialogue until it determines that sufficient information has been collected. The agent uses techniques such as slot filling, where it aims to fill predefined information "slots" based on the user's responses.

The interview agent's design implements the stakeholder integration principles highlighted by Liu et al. (2024), who emphasized the importance of gathering nuanced input from content experts at the beginning of the development process. This addresses the challenge identified in the needs assessment regarding the significant time required for collaboration between system designers and subject matter experts.

Planning Agent. After the information is gathered, the planning agent organizes it into a structured format. This agent focuses on fleshing out the user's intent, creating a high-level outline that will guide subsequent content development.

The planning agent operationalizes key principles of Cognitive Load Theory (Sweller, 2011) by organizing content into structured learning sequences that optimize cognitive resources. This directly addresses the challenge identified in the analysis phase that students, especially those with limited prior knowledge, struggled with cognitive overload when presented with complex data science concepts.

Writing Agent. The writing agent generates content based on the structured plan. It specializes in transforming the outline into detailed curriculum components, ensuring that the content aligns with the user's goals and is appropriate for the target audience.

The writing agent implements the contextual relevance principles identified in the analysis phase, addressing the finding that student engagement increased significantly when content incorporated examples from domains relevant to learners' interests. Its design is informed by Rotthoff et al.'s (2015) research on authentic contexts in problem-based learning.

Evaluation Agent. This agent reviews the content generated by the writing agent, assessing it for accuracy, coherence, and adherence to the original objectives. It checks for consistency with external knowledge sources and the QuantHub knowledge graph.

The evaluation agent's design is directly influenced by the quality assurance layer described by Woo et al. (2024), who demonstrated a 39.7% improvement in response accuracy through systematic validation against authoritative sources. This addresses the critical need for factual accuracy in data science education identified in the analysis phase.

Revising Agent. If the evaluation agent identifies issues or if the user provides feedback, the revising agent adjusts the content accordingly. This agent specializes in refining the output to meet the desired standards.

The revising agent implements the iterative refinement approach advocated by Dierickx et al. (2024), who found that human-AI collaboration in content validation significantly improves quality. This addresses the challenge identified in the needs assessment that manual revisions were time-consuming and created bottlenecks in the content development process.

Agent Collaboration and Workflow

The agents operate in a collaborative workflow, where each agent's output becomes the input for the next agent.

In the module outline development, the system employs an interaction between a student agent and a teacher agent (Fig. 1). The student agent, assigned the role of the target audience (e.g., an eighth grader), generates questions relevant to the curriculum topic. These questions reflect the curiosity and knowledge gaps of the intended learners.

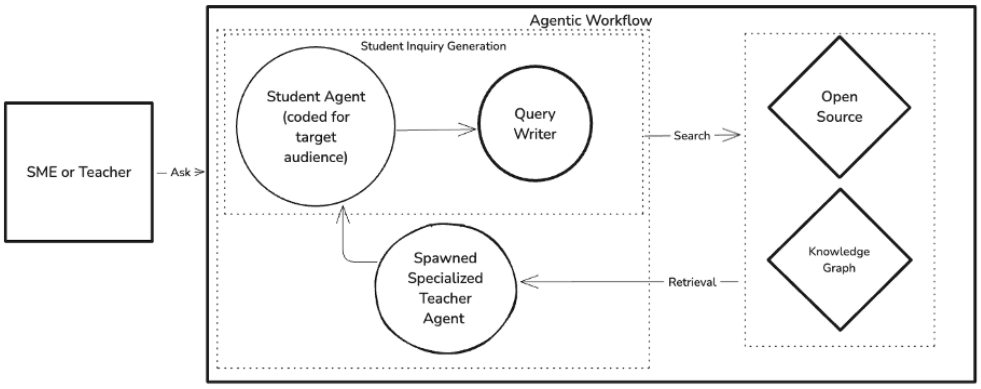
However, student-generated questions may not be optimally phrased for information retrieval due to their conversational nature or complexity. Therefore, a query transformation agent converts these questions into effective search queries. For example, a student's question like "How do plants eat if they don't have mouths?" might be transformed into "Photosynthesis process in plants."

The transformed queries are then used to search external knowledge sources, such as the web or external resources accessible to the agents. The search results provide factual information and citations, which are fed back to the teacher agent. The teacher agent uses this grounded information to generate accurate and contextually appropriate answers to the student's questions.

For example, the interview agent's collected information is passed to the planning agent, which then creates an outline that the writing agent uses to generate content. The evaluation agent reviews this content, and if revisions are needed, the revising agent adjusts before the content is finalized.

Figure 1

Agentic workflow within the Module Outline component of the curriculum builder.



This multi-agent approach improves the system's adaptability and reliability. By having specialized agents focus on specific tasks, the system reduces the variability in outputs and ensures that each component of the content is thoroughly developed and evaluated. This aligns with the findings of Singh et al. (2024), who demonstrated that multi-agent LLM systems outperform traditional linear interactions in terms of productivity and performance.

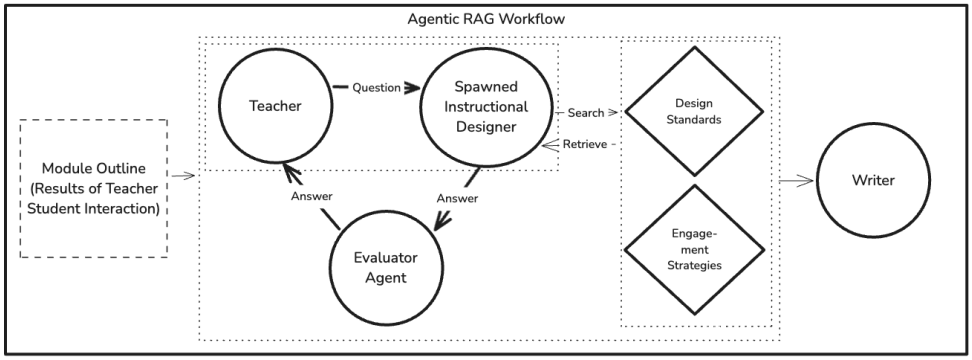
Incorporation of Retrieval-Augmented Generation (RAG)

To eliminate the risk of hallucinations and ensure that the AI output is grounded in research, the curriculum builder incorporates Retrieval-Augmented Generation (RAG) with corrective mechanisms.

Retrieval-Augmented Generation (RAG). The previous curriculum builder framework resulted in inconsistent content that demonstrated the LLMs desire to extrapolate, leverage misinformation, or provide nonsensical responses with a high level of confidence in the output. Additionally, when needing to support a curriculum that focused on emerging research and innovation, the model was limited due its out-of-date training data. To address these challenges, the new system employs RAG by feeding in frameworks, guidelines, and other external documentation that are well-researched and validated by subject matter experts. No longer does the system rely on two shot examples.

Figure 2

RAG within the development of the learning resource



Looking at the learning resource generation component of the system, RAG supports the incorporation of impactful instructional design engagement strategies (Fig. 2). Existing research shows that the use of specific engagement strategies supports improved learning. By providing the agent with those engagement strategies and contextual knowledge, the system leverages an external store when writing the learning resource. This strategy ensures the agent is not improvising engagement strategies but grounding its outputs in vetted learning science principles. By incorporating RAG, the system ensures that the

content generated is based on verified information, minimizing the risk of inaccuracies or hallucinations. This technique aligns with the research by Asai et al. (2023) and Yan et al. (2024), who highlight the effectiveness of RAG in grounding LLM outputs.

Corrective Mechanisms (CRAG). A self-correcting mechanism within the system is introduced through the evaluation agent. Within each component of the curriculum builder, this agent ensures curriculum content adheres to pedagogical principles by systematically reviewing the content against specific instructional design criteria. A key focus is verifying the alignment with principles of content organization and scaffolding, derived from foundational theories, to support effective learning.

1. Organizing from Simple to Complex: During the development of the Skill Outline, the Evaluation Agent checks whether tasks are sequenced logically, starting with foundational concepts and gradually introducing more complex ideas. For example, an introductory task might require learners to perform basic data cleaning, while subsequent tasks guide them through increasingly intricate processes such as feature engineering or predictive modeling. This approach ensures learners develop a robust understanding before tackling advanced concepts.
2. Managing Cognitive Load: During the development of the module outline, the agent evaluates concepts to confirm that they present information in manageable chunks. This involves ensuring that new content builds on existing knowledge and avoids overloading students with unrelated or excessive details. For instance, concepts are structured to introduce one new technique at a time, coupled with reinforcement exercises that applies that technique in a relevant context. The agent flags any concepts that present too much information at once or lack clear language, recommending adjustments to improve clarity and focus.
3. Fostering Conceptual Connections: A critical component of the agent's review is verifying that content encourages students to make connections between theoretical concepts and practical applications. For example, an exercise supporting statistical hypothesis testing might include scenarios that apply these concepts to real-world datasets, such as evaluating marketing campaigns or analyzing public health data. The agent checks for prompts that require learners to reflect on how prior tasks inform their current work, reinforcing the interconnected nature of the material.
4. Ensuring Contextual Relevance and Engagement: The agent uses information about intended learners gathered during the skill outline process to evaluate whether the content is designed to resonate with learners. By incorporating language and examples relevant to the intended learner, the curriculum aims to maintain learner interest and demonstrate the applicability of content. Content that does not meet provided standards are flagged for revision to better engage students and align with expected learner achievement.

By applying these principles, the Evaluation Agent ensures that the curriculum supports incremental learning, optimizes cognitive resources, and promotes meaningful connections between concepts, while maintaining high levels of relevance and engagement.

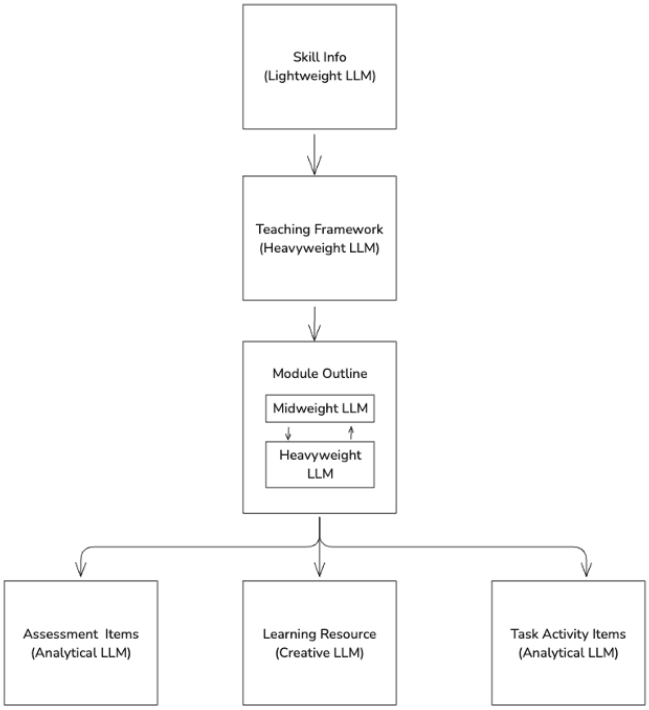
Selection and Integration of Different LLMs

The curriculum builder leverages different LLMs throughout the development process to capitalize on their unique strengths and specializations. This approach is informed by the research of Wang et al. (2024), which emphasizes the benefits of using multiple models for various task aspects.

Model Selection Based on Task Requirements. Foundational instructional design models (e.g., ADDIE & Gagne's Nine Events of Instruction) emphasize the importance of aligning instructional strategies and activities with well-defined learning objectives. This same principle is applied when aligning tools with intended content development. By selecting LLMs optimized for particular content creation tasks (Fig. 3), the builder ensures that the generated content directly supports the intended outcomes.

Figure 3

Example of alignment of LLMs within different components of the curriculum builder



Lightweight Models (e.g., GPT-4 Mini). Used for tasks that require speed and basic interactions, such as the initial interview agent's conversation with the user, these models are efficient in handling dynamic dialogues without needing extensive computational resources.

Midweight and Heavyweight Models (e.g., GPT-4). These models are employed for complex planning and content generation tasks where higher accuracy and depth are required. For example, the planning agent uses advanced models to create detailed outlines that form the backbone of the curriculum.

Creative Models (e.g., Claude). Utilized for generating stylized outputs and handling tasks that involve creativity and abstract thinking, the writing agent may use these models to produce engaging and pedagogically effective content.

Dynamic Switching and Prompt Adjustment. The system allows for dynamic switching between models based on performance needs without significant reconfiguration. When a new model becomes available or an existing model is updated, the system can integrate it by adjusting the prompts to align with the model's capabilities.

For example, if the GPT-4 Mini model is replaced with Claude for the interview agent, the prompts may need slight adjustments to account for the model's different conversational style. This flexibility ensures that the system remains adaptable and can continuously improve by leveraging advancements in LLM technology.

Agentic Workflow and Continuous Improvement

The use of an agentic workflow facilitates self-evaluation and multi-model evaluation, leading to continuous improvement within content development.

Self-Evaluation Mechanisms. Each agent in the system is designed to evaluate its own outputs and those of previous agents. For instance, the evaluation agent checks the content generated by the writing agent for accuracy and alignment with the initial objectives. If issues are identified, the revising agent makes necessary adjustments.

This self-evaluation mechanism ensures that errors are caught and corrected within the system before the content reaches the user. It enhances the reliability and validity of the final output.

Multi-Model Evaluation. By utilizing different LLMs for various tasks, the system can compare outputs from multiple models to identify inconsistencies or errors. For example, the same content might be generated by both GPT-4 and Claude, and the evaluation agent can assess which version better meets the quality standards.

This multi-model evaluation enables the system to combine the strengths of different models, leading to higher-quality content generation.

Incorporation of Personas and Styles

To enhance the tone and style of the content, the curriculum builder incorporates personas of well-known educators. Agents are assigned these personas to guide their output, ensuring that the content is engaging and appropriate for the target audience.

For example, an author profile might describe the teaching style of known subject matter expert and educator in the field, emphasizing clear explanations and patience. By abstracting these style characteristics, the system guides the writing agent to produce content that reflects these qualities without introducing personal anecdotes or content that deviates from the educational objectives.

Integration with Adaptive Learning Pathways

The architecture of the multi-agent LLM framework directly enables the granular content generation required for QuantHub's adaptive learning implementations. While the curriculum builder framework's primary function is the systematic development and validation of curriculum components, its significance extends to facilitating dynamic content orchestration. The system's integration of RAG and CRAG methodologies, coupled with its robust evaluation mechanisms, enables the generation of pedagogically diverse content modules that can be programmatically assembled based on learner analytics. This technical foundation—particularly the framework's capacity for verified, componentized content generation—is fundamental to implementing evidence-based adaptive learning at scale.

This integration approach implements the personalized learning principles emphasized by Jumaat and Tasir (2014), who demonstrated the importance of adaptive scaffolding in online learning environments. It addresses the finding from the needs assessment that students have diverse learning needs and require tailored support, particularly in complex domains like data science.

Testing and Refinement Phase: Iterative Implementation

Following the design-based research methodology, an iterative testing and refinement process was implemented to evaluate and enhance the AI curriculum builder. This phase focused on assessing the system's effectiveness in addressing the challenges identified in the analysis phase while gathering feedback for continuous improvement. Content developers and instructors implementing the curriculum were queried throughout this phase.

Testing Protocols and Evaluation Framework

The evaluation approach employed a mixed-methods framework designed to assess the system's impact on three key dimensions derived from the research questions:

1. Curriculum Quality and Accuracy: Evaluation criteria included content accuracy, scaffolding, contextual relevance, and alignment with learning objectives.
2. Developer and Teacher Experience: The system's impact was assessed on curriculum development efficiency, teacher effort, and adaptability to different educational contexts.

3. Student Learning Experience: Student persistence and confidence were evaluated when using the AI-generated curriculum materials.

For each dimension, specific indicators and metrics were developed, combining quantitative measures of content quality with qualitative assessments of user experience. This approach aligns with the design-based research principle of examining both practical effectiveness and theoretical contributions (Barab & Squire, 2004).

Data Collection Methods and Instruments

Multiple data collection methods were used to gain comprehensive insights into the system's performance:

1. Content Quality Evaluations: Subject matter experts assessed the quality of AI-generated curriculum components using a standardized rubric covering accuracy, pedagogical soundness, scaffolding quality, and contextual relevance.
2. Developer and Teacher Interviews: Weekly semi-structured interviews were conducted with content developers and teachers to gather insights on quality and implementation of generated content within existing course designs.
3. Student Usage Analytics: The learning platform collected data on student interaction patterns, including completion rates, progression through learning pathways, and performance on assessments. These analytics provided quantitative insights into student engagement and persistence.

This multi-faceted approach allowed for triangulation of findings across different stakeholder perspectives and data types, enhancing the validity of the evaluation without relying on automated system performance logs.

Results

The implementation of the AI curriculum builder showed significant impact across three key dimensions, directly addressing the research questions posed in this study. This section presents findings from the iterative testing and evaluation process, including quantitative data from student surveys and qualitative feedback from multiple stakeholders.

Addressing Research Question 1: Bridging the Gap Between Learning Theory and Practical Curriculum Design

The following focuses on how the AI-supported curriculum development framework operationalizes learning theories, demonstrating its effectiveness in translating theoretical principles into practical instructional materials.

Improved Content Quality and Pedagogical Alignment

The output of the AI curriculum builder demonstrated noteworthy improvements in consistent quality, valid output, and alignment to instructional design principles. The integration of RAG and agents specializing in designing based on evidence-based practices and pedagogical evaluation criteria resulted in curriculum materials that effectively implemented scaffolding principles, cognitive load management, and contextual relevance.

- Error rates relating to content alignment with skill intent dropped between 20-30%.
- Errors in design inconsistencies dropped more than 50%.
- One of the greatest improvements was in the development of distractors and explanations that enabled real-time feedback. Distractors were no longer obvious, but rather additional meaningful learning elements.

- The system was able to rapidly create context “wrappers” that allowed learners to engage with numerous applications, datasets, and workplace simulations. This aligns curriculum materials with real-world applications.

Addressing Research Question 2: Impact on Quality, Accuracy, and Contextual Relevance of Educational Content

This section presents findings related to how the integration of AI-driven workflows, specifically agentic models and RAG, influenced the quality, accuracy, and contextual relevance of the educational content generated.

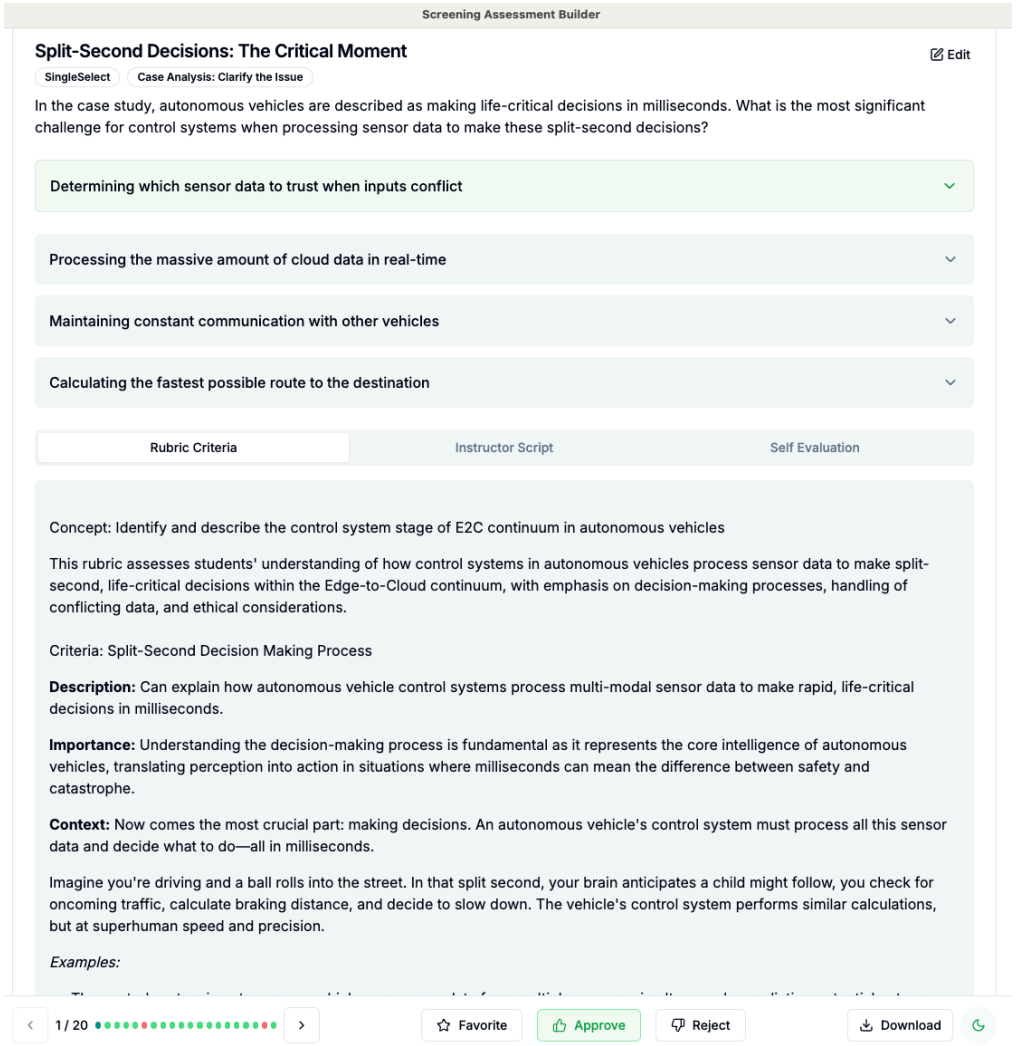
Enhanced Content Development Process and Accuracy

Members of the content development team reported consistent positive improvements in the curriculum building process. When evaluating the content development of an Applied Data Science course (curriculum builder output example shown below in Fig. 4), the incorporation of agent-based workflows featuring planning, writing, and evaluation cycles led to the following improvements.

- The frequency of misunderstandings dropped more than 20% in the initial skill outline phase, decreasing the need to rephrase initial commands.
- The subject matter review process, which previously created substantial bottlenecks, was streamlined through the new architecture and communication of agentic reasoning process.
- Content review for a single course dropped from 180 minutes to 45 minutes.
- AI accelerated the content development process, generating initial drafts and identifying areas for refinement, while SMEs provided essential oversight at key review points.
- Resource allocation for context-generation efforts was reduced by nearly 70%.

Figure 4

User interface for AI-generated content review. Rubric criteria are curated from an existing knowledge database.



Addressing Research Question 3: Affect on Teacher Effort, Student Engagement, Persistence, and Learner Confidence

Observed effects of the AI curriculum builder framework were captured for key stakeholders, including educators and students, focusing on their experiences and outcomes in the online learning environment.

Reduced Teacher Effort and Improved Educator Satisfaction

Initial beta testing revealed positive reception from instructors (n=6). They reported that the platform simplified the process of aligning newly generated materials with their established curricular goals.

- The improved structure of content deliverables allowed instructors to select and sequence activities that easily paired with their in-class instructional routines.
- The adaptability of the builder allowed quick adjustment of materials to seamlessly supplement core lessons and integrate new learning opportunities.
- Instructors particularly valued the system's ability to generate contextually relevant examples that resonated with their specific student populations. One instructor noted, "Being able to quickly customize examples to match my students'

interests made a huge difference in their engagement level. They're actually connecting these data concepts to their everyday lives".
- Another instructor highlighted the time-saving aspect: "Previously, I would spend hours creating context for abstract concepts. Now the system generates relevant examples that I can simply review and select, saving me significant preparation time".
- Instructors also indicated that the AI-generated materials matched appropriate levels of cognitive challenge, with reports that activities were neither overly simplistic nor overly complex.

Enhanced Student Engagement, Persistence, and Confidence

Survey data collected from students in two courses revealed significant improvements in self-reported understanding across multiple data science concepts after completing the AI-generated curriculum activities. The survey used a 1-5 scale of perceived understanding with 1 being no understanding and 5 as expert level understanding.

- In Course A (n=10), notable improvements were reported in students' understanding of Excel-based data analysis concepts. Results are indicated by gains.

- Organizing data in frequency tables in Excel (average increase of 1.7 points)
- Formatting frequency tables (average increase of 1.6 points)
- Creating charts for data analysis (average increase of 1.4 points)
- Making recommendations based on frequency tables (average increase of 1.3 points)

- Similarly, in the BUS 271 course (n=41), students identified significant gains in perceived understanding of more advanced time-series analysis concepts. Results are indicated as reported initial level of understanding and final level of understanding.

- Calculating and plotting moving averages (average increase from 2.2 to 3.9)
- Time-series analysis in Excel (average increase from 2.3 to 4.0)
- Building narratives around time-series data (average increase from 2.1 to 3.8)

- Prior to implementing the AI curriculum builder, completion rates for advanced analytical content were below 6%. In contrast, the new contextually relevant and appropriately scaffolded curriculum materials showed significantly higher engagement.
- Qualitative feedback from Course A students highlighted that "The order of the lessons made it easy to follow along, even when I didn't have prior experience with some of the Excel functions". They also noted, "I liked how the challenge related to real business situations. It helped me understand why we need to analyze data this way" , and "Having different examples of the same concept helped me see how to apply these skills in different situations".
- Course B students noted improvements in their confidence with data analysis, stating, "Before this challenge, I was intimidated by time-series analysis, but breaking it down into smaller tasks made it manageable" , "The way the materials explained the 'why' behind each step helped me understand the concepts, not just follow directions" , and "I appreciated how the challenge started simple and then gradually added complexity as I built confidence".
- In beta testing classrooms, students' persistence in activities did not experience the drop as seen with the previous content, suggesting that the improved scaffolding and contextual relevance effectively supported student engagement.

Further Research

Next steps include the further integration of a knowledge graph as a foundational tool in the curriculum builder architecture. The integration of a curated, systemically organized representation of concepts, learning outcomes, tasks, and standards will

further ensure the delivery of structured, transparent, and evidence-based instructional design. The graph will serve as the blueprint from which the curriculum will reliably generate content that is both contextually relevant and instructionally sound.

The heart of the current curriculum builder is the principle of aligning all content with clearly defined learning outcomes. By encoding these outcomes and their supporting concepts into the graph, the builder ensures that any content development the AI produces is directly traceable to the intended competencies. The outcome-focused alignment will continue to prevent the creation of extraneous or unanchored content, supporting the delivery of purposeful instruction and the system's ability to measure and communicate outcomes. This is key for supporting student metacognition.

Furthermore, the continued growth and integration of the graph supports the principle of data-informed iteration. Its structured relationships facilitate ongoing improvement of the curriculum builder, and subsequently, a learner's learning path. When SMEs, educators, and industry practitioners refine the AI output, these updates can be captured at a node level. The goal is to ensure the graph becomes a dynamic reference, continually evolving based on feedback captured by the curriculum builder and the QuantHub platform.

Ethical Considerations and Responsible Implementation

As the AI-supported curriculum development framework is advanced, several important ethical considerations must be addressed. First, there is the critical importance of transparency in AI-generated educational content. Instructors and students should understand when and how AI has contributed to curriculum materials, fostering an environment of informed use rather than opaque implementation. To this end, future development will include clear documentation and attribution systems that make AI contributions visible while maintaining instructional coherence.

Central to framework iteration is addressing representation and bias in content development. Educational materials have historically reflected and reinforced dominant cultural perspectives, often marginalizing diverse voices and experiences. The AI curriculum builder must actively counter these patterns rather than perpetuate them. Future research will prioritize developing comprehensive bias detection mechanisms that identify potentially exclusionary language, examples, or frameworks within generated content. This will include implementing representational heuristics that ensure diverse perspectives and cultural contexts are authentically incorporated throughout curriculum materials.

Conclusion

This research highlights the potential of AI in curriculum development, showcasing its ability to operationalize learning theories effectively. The AI-supported framework integrates RAG and agentic models to create scaffolded, contextually relevant, and adaptive learning materials, enhancing the alignment of curriculum design with pedagogical principles. Initial implementation results demonstrate improved engagement, persistence, and self-efficacy among learners. Moreover, the system reduces barriers for educators by streamlining content development and ensuring alignment with classroom standards. Future directions include expanding the integration of knowledge graphs to further personalize and adapt curriculum pathways, enhancing work-based learning opportunities, and exploring applications in professional certifications. These advancements represent a significant step forward in educational innovation, fostering both learner success and educator efficiency.

References

Aryal, S., Do, T., Heyojoo, B., Chataut, S., Gurung, B. D. S., Gadhamshetty, V., & Gnimpieba, E. (2024). Leveraging multi-AI agents for cross-domain knowledge discovery. arXiv preprint arXiv. https://doi.org/10.48550/arXiv.2404.08511

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences, 13*(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1

Chu, Z., Wang, Y., Zhu, F., Yu, L., Li, L., & Gu, J. (2024). Professional agents--Evolving large language models into autonomous experts with human-level competencies. arXiv preprint arXiv:2402.03628. https://doi.org/10.48550/arXiv.2402.03628

Cochran, K. (2024). Classroom observation notes.

Dierickx, L., Sirén-Heikel, S., & Lindén, C. G. (2024). Outsourcing, augmenting, or complicating: The dynamics of AI in fact-checking practices in the Nordics. *Emerging Media, 2*(3), 449-473. https://doi.org/10.1177/27523543241288846

Domingo, J. M. (2024). Curriculum planning, implementation, development and evaluation: strategies and challenges for modern education: A literature review. *Global Scientific Journal, 12*(11), 1089-1101. https://www.globalscientificjournal.com/researchpaper/_Curriculum_Planning_Implementation_Development_and_Evaluation_Strategies_and_Challenges_for_Modern_Education_A_Literature_Review_.pdf

Ertmer, P. A., & Newby, T. J. (2013). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance improvement quarterly, 26*(2), 43-71. https://doi.org/10.1002/piq.21143

Fries, L., Son, J. Y., Givvin, K. B., & Stigler, J. W. (2021). Practicing connections: A framework to guide instructional design for developing understanding in complex domains. *Educational Psychology Review, 33*(2), 739–762. https://doi.org/10.1007/s10648-020-09561-x

Gillani, N., Eynon, R., Chiabaut, C., & Finkel, K. (2023). Unpacking the "black box" of AI in education. *Educational Technology & Society, 26*(1), 99-111. https://doi.org/10.30191/ETS.202301_26(1).0008

Jumaat, N. F., & Tasir, Z. (2014, April). Instructional scaffolding in online learning environment: A meta-analysis. In 2014 international conference on teaching and learning in computing and engineering (pp. 74-77). IEEE. https://doi.org/10.1109/LaTiCE.2014.22

Kahl, S., Löffler, F., Maciol, M., Ridder, F., Schmitz, M., Spanagel, J., Wienkamp, J., Burgahn, C., & Schilling, M. (2024, July 10). Enhancing AI tutoring in robotics education: Evaluating the effect of retrieval-augmented generation and fine-tuning on large language models (Working Paper No. 9). Autonomous Intelligent Systems Group, University of Münster. https://www.uni-muenster.de/imperia/md/content/angewandteinformatik/aktivitaeten/publikationen/enhancing_ai_tutoring_in_robotics_education_-_2024.pdf

Li, X., Zhang, Y., & Malthouse, E. C. (2024). Large language model agentic approach to fact checking and fake news detection. In ECAI 2024 (pp. 2572-2579). IOS Press. https://doi.org/10.3233/FAIA240787

Liu, H., Das, A., Boltz, A., Zhou, D., Pinaroc, D., Lease, M., & Lee, M. K. (2024). Human-centered NLP Fact-checking: Co-Designing with Fact-checkers using Matchmaking for AI. https://doi.org/10.1145/3686962

Meng, N., Dong, Y., Roehrs, D., & Luan, L. (2023). Tackle implementation challenges in project-based learning: A survey study of PBL e-learning platforms. *Educational Technology Research and Development, 71*(3), 1179–1207. https://doi.org/10.1007/s11423-023-10202-7

Norman, M. K., & Lotrecchiano, G. R. (2021). Translating the learning sciences into practice: A primer for clinical and translational educators. *Journal of Clinical and Translational Science, 5*(1), e173. https://doi.org/10.1017/cts.2021.840

Reeves, T. C. (2006). Design research from a technology perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 52-66). Routledge. https://doi.org/10.4324/9780203088364-13

Reigeluth, C.M. and Stein, F.S. (1983) The elaboration theory of instruction. In C. Reigeluth (Ed.), *Instructional-design theories and models: An overview of their current status* (pp. 335-381). Lawrence Erlbaum Associates, Publishers.

Rotthoff, T., Schneider, M., Ritz-Timme, S., & Windolf, J. (2015). Theory in practice instead of theory versus practice--curricular design for task-based learning within a competency oriented curriculum. *GMS Zeitschrift fur medizinische Ausbildung, 32*(1), Doc4. https://doi.org/10.3205/zma000946

Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2024). Enhancing AI systems with agentic workflow patterns in large language models. In 2024 IEEE World AI IoT Congress (AIIoT) (pp. 527-532). IEEE. https://doi.org/10.36227/techrxiv.173092393.30216600/v1

Snyder, J. (2023). *QuantHub pilot analytics report*.

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (pp. 37–76). https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Wang, F., Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development, 53*, 5–23. https://doi.org/10.1007/BF02504682