LLM-Assisted, Iterative Curriculum Writing: A Human-Centered AI Approach in Finnish Higher Education

Leo Huovinen
Metropolia University
of Applied Sciences
Helsinki, Finland
leoeinari.huovinen
@metropolia.fi

Mika Hämäläinen
Metropolia University
of Applied Sciences
Helsinki, Finland
mika.hamalainen
@metropolia.fi

Abstract

This paper presents a Large Language Model (LLM)-based system designed to support curriculum development, iteratively refined through extensive user testing and deployed within a major Finnish higher education institution over the past two years. Distinct from typical content generation tools, our system facilitates iterative human-AI collaboration by providing structured suggestions and analyzing course descriptions for alignment with institutional goals, accreditation requirements, and competency frameworks. We investigate how such a tool can reduce educators' cognitive load while preserving human expertise, detailing the system's technical architecture and iterative development grounded in a human-centered design approach. This involved prototyping, workshops, and user testing with curriculum coordinators and faculty across diverse departments. We present detailed findings, including quantitative metrics, qualitative feedback, and user quotes, demonstrating the system's evolving reception and potential to support complex educational planning tasks.

1 Introduction

Curriculum development in higher education presents a significant challenge, demanding alignment with diverse stakeholder needs, established competency frameworks, and stringent quality-assurance standards (Barnett and Coate, 2005; Knight, 2001). Educators face increasing pressure to design curricula that satisfy institutional mandates and accreditation criteria while catering to the evolving requirements of diverse student populations (Teixeira et al., 2019; Oliver and Hyun, 2011). This complex task often results in considerable cognitive load, compounded by fragmented information systems and administrative hurdles (Woelert, 2023).

While artificial-intelligence (AI) tools for writing assistance have advanced rapidly (Strobl et al., 2021), generic AI systems often lack the specificity required for effective curriculum development. Key aspects such as alignment with competency frameworks, nuanced assessment design, and adherence to regulatory compliance are frequently inadequately addressed (Zawacki-Richter et al., 2019). Early applications of language models in education predominantly focused on content generation, rather than supporting the inherently iterative and collaborative nature of curriculum writing (Roll and Wylie, 2016; Huang et al., 2023), often failing to alleviate the core challenges faced by educators.

In response to these limitations, we developed an LLM-based curriculum-development system designed as an interactive collaborator. Our approach emphasises maintaining human expertise and agency throughout the writing process, shifting the focus from mere automation to synergistic human-AI partnership (Holstein et al., 2019; Kamar, 2016; Wilson and Daugherty, 2018). This system has been iteratively developed and tested over 18 months at a multi-disciplinary university of applied sciences in Finland. Figure 1 shows the deployed system in active use, analyzing a nursing science master's degree curriculum against UN SDGs, illustrating the practical application of our iterative design process.

This study explores several critical aspects through the lens of our development and deployment experience. We investigate how an LLM-assisted tool can reduce the cognitive load on educators during curriculum development, presenting evidence from user testing. We examine how such a tool can effectively support the alignment of curriculum content with institutional goals, accreditation standards (e.g. UN Sustainable Development Goals), and competency frameworks, reporting on user experiences with these features. Additionally, we consider how the system design, informed by user feedback, accommodates varying levels of AI

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1002–1010
July 31 - August 1, 2025

Figure 1: The deployed curriculum writing tool interface showing analysis of a Master's Degree Programme in Development and Leadership of Nursing degree program. The left panel displays the curriculum structure with courses organized by specialization tracks. The center shows automated LLM analysis mapping course content to UN Sustainable Development Goals through bar charts and pie visualization. The right panel provides detailed SDG alignment feedback, demonstrating the system's capability to analyze curriculum content against institutional frameworks and provide structured guidance to educators.

literacy among faculty members. Finally, we detail how an iterative, human-centred design process, combining user tests and workshops, effectively refined the tool for practical integration into institutional workflows.

2 Related Work

2.1 Curriculum Development Challenges and Educator Needs

Curriculum development is a cornerstone of educational practice, demanding alignment across diverse requirements such as institutional goals, pedagogical principles, accreditation standards, and learner needs (Barnett and Coate, 2005; Knight, 2001). Educators tasked with this complex endeavour often face significant cognitive load (Sweller, 1988). Existing digital tools frequently fall short, hampered by usability issues, poor integration, and failure to streamline workflows (Woelert, 2023; Fernández-Cerero et al., 2024). This can lead to frustration among educators who find such tools increase administrative burden rather than reduce it (Blaich and Wise, 2018; Sjöberg and Lilja, 2019; Duarte and Vardasca, 2023). Introducing AI therefore necessitates building trust; educator adoption hinges on understanding how AI functions and perceiving it as a supportive partner that complements their expertise (Nazaretsky et al., 2022). Addressing these usability, workflow, and trust challenges for educators is paramount.

2.2 NLP Applications for Curriculum Analysis and Related Tasks

Applying Natural Language Processing in education has often involved building specialised pipelines for narrow analytical tasks, frequently requiring substantial feature engineering. Areas such as automated essay scoring, grammatical-error correction (Bryant et al., 2019), and readability assessment (Aluisio et al., 2010) have seen dedicated development, yet applying NLP effectively to *curriculum development* presents unique challenges.

A central task is ensuring semantic alignment between components like learning outcomes, course content, and assessments, and verifying coverage of external competency frameworks. Early NLP approaches tackled this via greedy similarity metrics (Rus and Lintean, 2012) or by constructing educational knowledge graphs through concept linkage (Dang et al., 2021). Analysing curriculum structure is another key requirement. Techniques for extracting prerequisite relations increasingly model course networks as graphs; recent work employs heterogeneous graph neural networks to infer prerequisite links from course-sequence data (Roy et al., 2019).

While effective for specific goals, these examples illustrate a trend towards fragmentation: distinct models, feature sets, and separate tools were developed for semantic similarity, structural relations, quality attributes, or content generation. Supporting the holistic process of curriculum writing; integrating multiple analyses and feedback remains difficult with such pipeline-based approaches.

2.3 Human-Centred Design: Bridging NLP Power and Educator Usability

NLP's analytical power only translates into impact when integrated via human-centred design (HCD). Educational settings involve diverse users (Gulbahar, 2008), and the cognitive load of complex tools is a major barrier (Sweller, 1988; Paas et al., 2003). Iterative HCD-workshops, prototyping and usability testing is essential for creating educational technology that educators find intuitive, trustworthy, and supportive (Druin, 2002; Quintana et al., 2004). For AI tools, transparency and user control are vital (Nazaretsky et al., 2022). Designing AI systems as collaborative partners that augment educator capabilities (Holstein et al., 2019) is therefore central to our methodology.

2.4 Prompt Engineering for Curriculum Development

Large foundation models such as PaLM-2[1] offer a shift away from fragmented pipelines. Effectively using these general-purpose models for specialised educational tasks relies on prompt engineering. While fine-tuning adapts models (Touvron et al., 2023), carefully crafted prompts can steer an LLM (Brown et al., 2020). Chain-of-thought prompting encourages structured reasoning suitable for alignment checks (Wei et al., 2022). Prompting for structured output (e.g. JSON) permits automatic parsing and presentation to educators, bridging the gap between raw LLM output and usable assistance (Ouyang et al., 2022).

2.5 Our Contribution: An Integrated, Human-Centred LLM Application

We present an LLM-assisted system co-designed as a collaborative partner for curriculum writing. Our contribution lies in a rigorous HCD process and a system architecture that prioritises educator usability, cognitive-load reduction, and integrated support for curriculum alignment. We move beyond fragmented individual NLP tools, such as semantic similarity analysis (Rus and Lintean, 2012; Dang et al., 2021), prerequisite extraction (Roy et al., 2019), readability (Aluisio et al., 2010) and error detection (Leacock et al., 2014), to leverage a single foundation model (PaLM-2). Carefully engineered prompts and a transparent UI provide unified support for alignment, quality checks, and structured suggestions.

3 Methodology

Our methodology employed a human-centered design (HCD) approach over an 18-month period, focusing on iterative development informed by continuous user feedback from the target end-users within a major Finnish university of applied sciences.

3.1 User-Centered Development Activities

We engaged curriculum coordinators and faculty members from diverse disciplines including Healthcare, Architecture, Therapeutic Studies, Engineering, and Business. Our development activities involved several key interactions. One-on-one usability testing occurred in January-February 2024 with 5 curriculum coordinators using early prototypes. These sessions involved participants performing domain-specific tasks, such as analyzing their 2024 curriculum against UN SDGs, institutional goals, and workplace requirements, while using a think-aloud protocol. Sessions were observed and recorded for qualitative analysis. Additionally, two major workshops were conducted as qualitative feedback sessions. The first, on June 6th, 2024, brought together 12 participants from five faculties for a demo presentation followed by hands-on testing and group discussions with casual Q&A. The second workshop, held on November 8, 2024, in a hybrid format, included 14 participants (both previous and new users) and followed a similar format of demo presentation, testing, and discussion, focusing on gathering requirements for features and integration priorities. Separately from these workshops, the demo tool was made publicly available via the institution's internal staff website, allowing independent access and usage. Throughout this process, feedback was collected via multiple

[1] https://ai.google/discover/palm2/

channels: interview notes, observation logs during testing, workshop discussions, and open-ended survey questions provided qualitative data from the interactive workshop sessions, while quantitative data was collected through a System Usability Scale (SUS)-inspired online feedback form completed by curriculum coordinators and faculty who accessed and used the demo tool independently via the internal website. This multi-faceted approach allowed us to identify usability requirements, cognitive load points, and evolving user needs, particularly regarding varying levels of AI literacy and integration with existing workflows.

3.2 Technical Architecture and System Implementation

The system was developed with a focus on modularity, scalability, and integration capabilities, employing a specific technical stack. The backend was implemented in Python[2] using the Flask[3] framework, deployed with uWSGI[4] behind an Nginx[5] reverse proxy on Debian/Ubuntu[6] Linux servers hosted within the institution's infrastructure; basic HTTP authentication via Nginx provided access control. For data persistence, MongoDB[7] (v. 7.0.1, initially with access control disabled during early development) served as the NoSQL database, storing curriculum data in a queryable format from the organization's curriculum database, organized by year (e.g., 2023, 2024, 2025), and recording document update timestamps. The frontend was a single-page application (SPA) built with React[8], utilizing React's Context API and hooks for state management, and communicating with the backend via RESTful API calls secured with CORS configuration. AI integration leveraged Google's Vertex AI[9] platform, specifically accessing the multilingual PaLM-2 foundation model through predefined prompt templates engineered to request structured JSON output, enabling reliable parsing and presentation of targeted feedback within the user interface in both Finnish and English. Security and logging included Nginx handling HTTPS encryption via SSL/TLS certificates and maintaining basic Nginx access logs with a 14-day rotation; application-level user interaction logging was minimal to prioritize privacy, which limited retrospective usage analysis but showed approximately 5 unique IP addresses accessing the API during a representative 14-day testing period. This architecture allowed for iterative updates to components like the LLM or UI while maintaining core functionality.

[2] https://www.python.org/
[3] https://flask.palletsprojects.com/
[4] https://uwsgi-docs.readthedocs.io/
[5] https://nginx.org/
[6] Debian: https://www.debian.org/, Ubuntu: https://ubuntu.com/
[7] https://www.mongodb.com/
[8] https://react.dev/
[9] https://cloud.google.com/vertex-ai

3.3 Iterative, Human-Centered Design Process

The 18-month development cycle unfolded following HCD principles across three main stages. The first stage focused on initial prototyping and testing, involving one-on-one tests (Jan-Feb 2024) for core concept validation and identifying fundamental usability issues, with feedback primarily concerning navigation and initial orientation. The second stage incorporated this feedback into a more robust prototype presented at the June 2024 workshop; this phase highlighted user needs for clearer guidance and workflow structuring to reduce cognitive load. The third stage addressed feedback from the first workshop and gathered requirements for more sophisticated functionality during the November 2024 workshop. In this final stage, user requests shifted towards advanced capabilities such as integration with the Peppi student information system, import features for existing drafts, quality control mechanisms, and enhanced multilingual and domain-specific support. Throughout this entire process, both qualitative and quantitative user feedback continuously informed design adjustments, feature prioritization, and refinement of the AI interaction model.

4 Results and Evolution of User Feedback

The iterative HCD process yielded rich insights into user needs and the system's effectiveness, revealing a clear evolution in feedback as the tool matured and users gained familiarity.

4.1 Initial Usability Testing (Jan-Feb 2024)

One-on-one sessions with 5 curriculum coordinators using early prototypes highlighted fundamental usability challenges and cognitive load concerns.

Orientation and Guidance: Users frequently expressed confusion upon first use:

> "There is no clarification here, I wouldn't know what this is. It wouldn't hurt to have a tool guide."

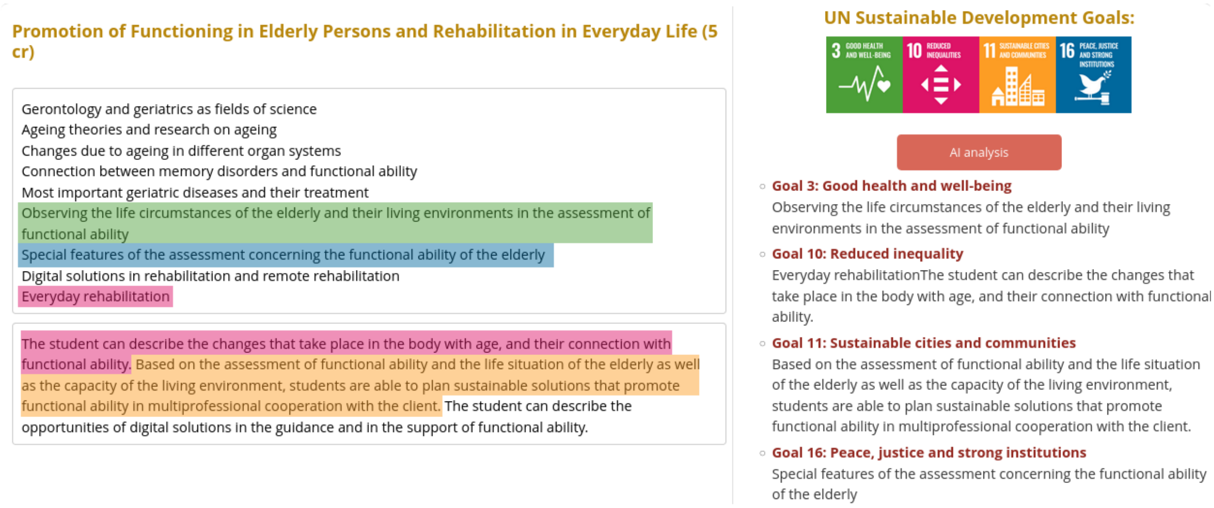

Figure 2: Screenshot of the prototype interface. The left panel shows course topics and learning outcomes; colour highlights indicate segments automatically matched by the LLM. The right panel lists the UN Sustainable Development Goals (SDGs) with the corresponding curriculum fragments, illustrating the tool's alignment-analysis feature.

> "Computer tools are not my favorite thing to deal with. I would like a guaranteed clarification of what I need to do with just a glance..."

Several users failed to notice key interface elements like year selection buttons, indicating issues with visual hierarchy. As one coordinator noted, "I didn't even see the year button. Maybe it could be larger or pointed out to me with some kind of guide?".

Cognitive Load from System Fragmentation: Users expressed frustration with managing information across multiple existing institutional tools.

> "There is a huge amount of information in different databases and tools in this house, but it is always difficult to find... it always takes a long time to find what I need."

> "It is frustrating to fill all kinds of online sticky notes with these goals, and then not have the time or coordination to apply these goals anywhere within the actual teaching."

Another user lamented the typical workflow: "Every year we have to jump between Peppi, Excel sheets, Teams files, and meeting notes. It's exhausting and error-prone." This highlighted the need for better integration and workflow streamlining, especially among faculties such as healthcare, which are burdened by need to align with several additional national and institutional standards (**?**).

Observational Data: During these sessions, observers noted consistent patterns: initial hesitation, significant time spent exploring menus before starting tasks, and a tendency to restart tasks rather than troubleshoot upon encountering errors. However, users showed visible positive reactions (surprise, verbal approval) when seeing the structured analysis generated by the AI, recognizing its potential time-saving benefits.

4.2 Workshop Feedback Evolution

First Workshop (June 6, 2024): With 12 participants from 5 faculties, this session confirmed persistent issues with navigation and cognitive load. The primary feedback emphasized the need for much clearer step-by-step guidance within the tool to structure the writing process itself, moving beyond simple text generation towards active workflow support.

Later Workshop (November 8, 2024): This session with 14 participants (hybrid format) showed a distinct shift. Having addressed basic usability, requests focused on advanced functionalities. Key demands emerged, such as a strong need for importing existing draft curriculum texts for LLM analysis and revision. Repeated requests were made for seamless integration with the institutional Peppi student information system to avoid redundant data entry. There was also a clear need expressed for better handling of discipline-specific terminology and requirements, particularly voiced by Healthcare and Engineering faculty coordinators; one noted, "Within our healthcare domain... there are specific training hours and certain areas of expertise that the students must meet," while another expressed concern about nuance: "AI is not able to detect all the 'weak signals'... I can recognize teaching-related problems... that I'm not sure AI would catch". Finally, users expressed a desire for features ensuring quality control and alignment with institutional standards and competency frameworks.

Participants consistently reiterated the importance of the AI acting as a collaborator. One from healthcare faculty stated, "I hope we can spend the most time on industry-specific goals... I'd like the easiest available tool for these general overhead tasks... We can then focus on our own expertise."

Experiences with Generic LLMs: Users who had tried generic tools like ChatGPT for curriculum tasks reported difficulties:

> "I have used AI (ChatGPT)... I tried to ask the AI to integrate the principles of sustainable development into this course, but what came out was difficult to use. I had to ask over and over to get the result I need."

This highlighted the value of our tool's structured approach and tailored prompts.

4.3 Quantitative Evaluation

Following the qualitative feedback, a modified SUS-style questionnaire focused on problem reporting was administered to participants familiar with the refined system. Due to its targeted nature, the sample size was small (n=4). The results (Table 1) indicate strong perceived utility for finding information (M=4.2) and content review (M=4.1), and high potential transferability (M=4.3). Interface usability (M=3.5) and learning curve (M=2.3, inverse scale) showed higher variability (SD=1.1, 1.2 respectively), supporting qualitative feedback about differing experiences based on user background and the need for continued ease-of-use improvements. Technical reliability was rated reasonably well (M=2.0, inverse scale).

Users also requested clearer visual feedback on standards coverage. Figure 2 illustrates the final UI that emerged from these iterations, showing colour-coded curriculum fragments mapped to specific SDGs.

5 Discussion

Our study demonstrates the potential for a carefully designed LLM-assisted tool, developed through an iterative, human-centered process, to effectively support the complex task of curriculum development in higher education. The detailed results from user testing (Section 4) provide concrete evidence addressing our core research questions.

The significant reduction in cognitive load was a key goal. Initial feedback highlighting confusion ("no clarification here...") and frustration with fragmented systems ("jump between Peppi, Excel sheets...") directly informed design iterations focused on providing clearer guidance and structured workflows. While full integration remains a challenge, the positive reception of the AI's structured analysis capabilities and the high rating for "Utility for content review" (M=4.1, see Table 1) suggest the tool successfully offloads some analytical burden. The shift in later feedback towards requesting deeper integration further indicates users perceived the tool's potential to streamline their work.

The system's ability to support alignment with institutional goals, accreditation standards (like UN SDGs), and competency frameworks was validated by user tasks during testing and the specific requests for enhanced quality control features in later workshops. The technical choice to use PaLM-2 via Vertex AI with structured JSON output proved crucial, enabling the system to provide targeted analysis rather than generic text, addressing the shortcomings users experienced with tools like ChatGPT ("had to ask over and over...").

Preserving human expertise was paramount. User quotes consistently emphasized the need for the AI to be a collaborator, handling "general overhead tasks" so educators could "focus on our own expertise" and address domain-specific nuances or "weak signals". The iterative design allowed us to balance automated assistance with user control, ensuring the tool augmented rather than replaced pedagogical judgment (Holstein et al., 2019; Kamar, 2016).

Accommodating varying AI literacy was implicitly addressed through the iterative process. Initial focus on fundamental usability ("guaranteed clarification... with just a glance") catered to less tech-savvy users, while later feature requests (import, advanced analysis) reflected the growing confidence and demands of users becoming more familiar with AI capabilities. The quantitative results showing

Table 1: Teacher Feedback (5-point Likert scale, n=4)
Items use inverse scale where lower scores indicate better performance.

Curriculum design task helpfulness criteria	Mean Score	Std. Dev.
Finding up-to-date degree info	4.2	0.8
Interface usability	3.5	1.1
Clarity of instructions	3.8	0.9
Learning curve (1=Easy, 5=Hard)*	2.3	1.2
Utility for content review	4.1	0.7
Technical reliability (1=Reliable, 5=Unreliable)*	2.0	0.7
Interface readability	3.9	0.5
Output and transferability of results	4.3	0.6

variance in usability and learning curve scores (Table 1) reinforce the need for continued attention to accessibility for all users.

The technical architecture (Section 3) supported this iterative development. The modular Flask/React stack and the use of a managed AI service (Vertex AI) facilitated relatively rapid prototyping and incorporation of feedback. The specific backend choices (Python, MongoDB, uWSGI, Nginx on Linux) represent a pragmatic and common stack for such institutional tools.

Challenges remain, particularly the significant technical and administrative hurdles of deep integration with complex systems (Brown et al., 2015; Sholeh et al., 2025). Supporting highly specialized disciplinary nuances and extending robust support for specific Finnish academic language or potentially underrepresented languages require ongoing effort. However, the positive trajectory of user feedback validates the HCD methodology and the potential of specialized LLM tools for complex educational planning.

6 Limitations and Future Work

While the HCD process yielded valuable insights and a functional tool, several limitations exist. The 18-month development timeline, driven by institutional curriculum renewal cycles, meant full integration with systems like Peppi was not achieved, limiting immediate efficiency gains highlighted as desirable by users ("jump between Peppi, Excel sheets..."). The automated analysis criteria were initially based on available institutional frameworks and UN SDGs; refining these for deeper discipline-specific requirements needs further work, as noted by users concerned about healthcare standards or engineering "weak signals."

The quantitative evaluation presented (Table 1) is based on a small sample size (n=4), limiting generalizability; it primarily served to corroborate qualitative findings during the iterative process. While the PaLM-2 model offers multilingual capabilities, dedicated fine-tuning or prompt optimization for specific Finnish academic contexts or other languages (e.g., Sámi languages) was beyond the scope of this phase.

While our approach relies on prompt engineering to adapt the general-purpose PaLM-2 model for curriculum development tasks, this may be insufficient for optimal performance in highly specialized domains. Effective prompts can improve output quality and structure, but cannot fully compensate for potential gaps in domain-specific training data or the nuanced understanding that dedicated fine-tuning or domain adaptation might provide. For instance, highly technical healthcare curriculum requirements or engineering accreditation standards may benefit from models specifically trained on educational content within those disciplines. Future work should explore whether fine-tuning approaches or domain-adapted models would significantly improve alignment accuracy and reduce the need for extensive prompt iteration.

The evaluation focused heavily on usability and perceived usefulness during development. Longitudinal studies are crucial to assess the tool's sustained impact on actual curriculum quality, alignment consistency across departments, and measurable changes in educator workload and satisfaction over time. Systematically evaluating effectiveness across a wider range of disciplines is also necessary. The minimal application-level logging, while prioritizing privacy, restricts retrospective analysis of feature adoption and user pathways.

Future work will prioritize tackling the Peppi integration challenge to enhance workflow automation. We plan to collaborate further with faculty to refine domain-specific analysis capabilities and expand language support. Exploring mechanisms for secure sharing of curriculum components or best practices across departments or potentially institutions represents another avenue. Rigorous, long-term evaluations measuring impact on curriculum outcomes and educator efficiency are essential next steps to guide continued refinement and demonstrate long-term value. Improving backend logging for anonymized usage patterns, while respecting privacy, would also aid future development.

7 Conclusion

This paper detailed the design, development, and user-centered evaluation of an LLM-assisted curriculum writing tool deployed at a major Finnish university of applied sciences. Through an 18-month iterative HCD process involving extensive user testing with curriculum coordinators and faculty, we created a system intended as a collaborative partner, aiming to reduce cognitive load and enhance alignment with standards, rather than simply automating writing. We presented specific technical details of the system (Python/Flask backend, React frontend, MongoDB, Vertex AI/PaLM-2 integration) and rich qualitative and quantitative data from user tests and workshops. The evolution of user feedback, from initial usability concerns ("no clarification here...") to demands for advanced features like Peppi integration and sophisticated analysis, strongly validates the iterative methodology. Our findings indicate that specialized LLM tools, co-designed with educators and focused on structured assistance, can effectively support complex educational planning tasks while preserving human expertise. While challenges in integration and domain specificity persist, this work offers a practical case study and valuable insights into developing human-centered AI solutions for higher education workflows.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the Workshop on NLP for Educational Applications*, pages 1–9.

Ronald Barnett and Kelly Coate. 2005. *Engaging the Curriculum in Higher Education*. Open University Press, Maidenhead, UK.

Charles Blaich and Kathleen Wise. 2018. Scope, cost, or speed: Choose two—the iron triangle of assessment. *Change: The Magazine of Higher Learning*, 50(3-4):73–77.

Malcolm Brown, Joanne Dehoney, and Nancy Millichap. 2015. The next generation digital learning environment: A report on research. Technical report, EDUCAUSE Learning Initiative.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The bea-2019 shared task on grammatical error correction. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75.

Fu-Rong Dang, Jin-Tao Tang, Kun-Yuan Pang, Ting Wang, Sha-Sha Li, and Xiao Li. 2021. Constructing an educational knowledge graph with concepts linked to wikipedia. *Journal of Computer Science and Technology*, 36(5):1200–1211.

Allison Druin. 2002. The role of children in the design of new technology. *Behaviour & information technology*, 21(1):1–25.

Nelson Duarte and Ricardo Vardasca. 2023. Literature review of accreditation systems in higher education. *Education Sciences*, 13(6):582.

José Fernández-Cerero, Julio Cabero-Almenara, and Marta Montenegro-Rueda. 2024. Technological tools in higher education: A qualitative analysis from the perspective of students with disabilities. *Education Sciences*, 14(3):310.

Yasemin Gulbahar. 2008. ICT usage in higher education: A case study on preservice teachers and instructors. *The Turkish Online Journal of Educational Technology-TOJET*, 7(1):32–37.

Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In *Artificial Intelligence in Education: 19th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I*, pages 157–168. Springer International Publishing.

Xinyi Huang, Di Zou, G. Cheng, X. Chen, and H. Xie. 2023. Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society*, 26(1):112–131.

Ece Kamar. 2016. Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4070–4073.

Peter T. Knight. 2001. Complexity and curriculum: A process approach to curriculum-making. *Teaching in Higher Education*, 6(3):369–381.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*, second edition. Synthesis Lectures on Human Language Technologies. Springer.

Tanya Nazaretsky, Moriah Ariely, Mutlu Cukurova, and Giora Alexandron. 2022. Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, 53(4):914–931.

Beverley Oliver and Eunsook Hyun. 2011. Comprehensive curriculum reform in higher education: collaborative engagement of faculty and administrators. *Journal of Case Studies in Education*, 2:1–20.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Fred Paas, Alexander Renkl, and John Sweller. 2003. Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, 38(1):1–4.

Chris Quintana, Brian J Reiser, Elizabeth A Davis, Joseph Krajcik, Eric Fretz, Ravit Golan Duncan, Eleni Kyza, Daniel Edelson, and Elliot Soloway. 2004. A scaffolding design framework for software to support science inquiry. *The journal of the learning sciences*, 13(3):337–386.

Ido Roll and Ruth Wylie. 2016. Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26(2):582–599.

Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. 2019. Inferring concept prerequisite relations from online educational resources. In *Proceedings of the AAAI Conference on Innovative Applications of Artificial Intelligence*, pages 9576–9583.

Varvara Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on NLP for Educational Applications*, pages 128–136.

Moch. Badrus Sholeh, Renita Fauziah Samodra, and Aris Puji Widodo. 2025. Benefits and challenges of erp implementation in higher education institutions: A systematic literature review. *Jurnal Sistem Informasi Bisnis*, 15(1):21–33.

Johan Sjöberg and Petter Lilja. 2019. University teachers' ambivalence about the digital transformation of higher education. *International Journal of Learning, Teaching and Educational Research*, 18(13):133–149.

Carola Strobl, Emilie Ailhaud, Kalliopi Benetos, Ann Devitt, Otto Kruse, Antje Proske, and Christian Rapp. 2021. Digital support for academic writing: A review of technologies and pedagogies. *Computers Education*, 166:104173.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

António Teixeira, Tony Bates, and José Mota. 2019. What future(s) for distance education universities? Towards an open network-based approach. *The International Review of Research in Open and Distributed Learning*, 20(4):1–19.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

H. James Wilson and Paul R. Daugherty. 2018. Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, 96(4):114–123.

Peter Woelert. 2023. Administrative burden in higher education institutions: A conceptualisation and a research agenda. *Journal of Higher Education Policy and Management*, 45(4):409–422.

Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):1–27.