Leo Huovinen

Assessing Usability of Large Language Models in Education

Metropolia University of Applied Sciences

Bachelor of Engineering

Information and Communication Technology

Bachelor's Thesis

23rd February 2024

Abstract

Author:	Leo Huovinen
Title:	Assessing Usability of Large Language Models in Education
Number of Pages:	27 pages
Date:	23rd February 2024
Degree:	Bachelor of Engineering
Degree Programme:	Information and Communication Technology
Professional Major:	Information and Communication Technology
Supervisors:	Mika Hämäläinen, Project Manager Janne Salonen, Head of School

The aim of the final year project was to investigate the potential of utilizing Large Language Models (LLMs) in automating and enhancing the process of curriculum development. The study sought to investigate how these models could streamline the laborious tasks involved in curriculum planning.

The project was carried out through the development and implementation of the LLM-based web application, followed by usability testing using the cognitive walkthrough approach and user feedback sessions. Test users were presented with the tool and tasked with evaluating its effectiveness, clarity, and overall usability in the context of curriculum planning. The study aimed to understand user interactions with the tool and identify areas for improvement to enhance its usability.

The results of this study show that while the LLM-based tool has the potential to streamline certain aspects of curriculum planning, there are key areas that require attention. Feedback from users highlighted the importance of clear guidance, integration of tasks into a centralized tool, and the need for human input in verifying and refining generated content, to ensure accurate and meaningful goals in curriculum design. These findings will guide the further development and refinement of the application.

Keywords: LLM, NLP, AI, Usability, Education, Curriculums

Tiivistelmä

Tekijä:	Leo Huovinen
Otsikko:	Suurten kielimallien hyödynnettävyyden arviointi opetuksessa
Sivumäärä:	27 sivua
Aika:	23.2.2024
Tutkinto:	Insinööri (AMK)
Tutkinto-ohjelma:	Tieto- ja viestintätekniikan tutkinto-ohjelma
Ammatillinen pääaine:	Tieto- ja viestintätekniikan tutkinto-ohjelma
Ohjaajat:	Mika Hämäläinen, Projektipäälikkö Janne Salonen, Osaamisaluejohtaja

Tämän opinnäytetyön tavoitteena oli tutkia suurten kielimallien (Large Language Models, LLM) potentiaalia opetussuunnitelmien kehittämistyön automatisoinnissa ja tehostamisessa. Tutkimus pyrki selvittämään, miten nämä mallit voisivat virtaviivaistaa työläitä tehtäviä, jotka liittyvät opetussuunnitelman suunnitteluun.

Opinnäytetyö toteutettiin kehittämällä ja toteuttamalla LLM-pohjainen webapplikaatio, jonka käytettävyyttä testattiin kognitiivisen kävelyn lähestymistavan ja käyttäjäpalautteen kautta. Testikäyttäjille esiteltiin työkalu ja heidät pyydettiin arvioimaan sen tehokkuutta, selkeyttä ja yleistä käytettävyyttä opetussuunnitelmien suunnittelukontekstissa. Tutkimuksen tavoitteena oli ymmärtää käyttäjien vuorovaikutusta työkalun kanssa ja tunnistaa käytettävyyteen liittyviä kehityskohteita.

Tutkimuksen tulokset osoittavat, että vaikka LLM-pohjaisella työkalulla on potentiaalia vähentää työläitä tehtäviä opetussuunnitelman suunnittelussa, on vielä olemassa huomiota vaativia osa-alueita. Käyttäjäpalautteiden perusteella korostui selkeän ohjeistuksen, tehtävien integroimisen keskitettyyn työkaluun sekä ihmisen panoksen tarve tuotetun sisällön tarkistamisessa ja hienosäätämisessä. Nämä havainnot ohjaavat työkalun jatkokehitystä ja hienosäätöä.

Avainsanat: LLM, NLP, AI, tekoäly, käytettävyys, opetus, opintosuunnitelmat

Contents

List of Abbreviations

LLM:	Large Language Models
NLP:	Natural Language Processing
UN:	the United Nations
SDG:	United Nation’s Sustainable Development Goals
GPT:	Generative Pre-trained Transformer

1 Introduction

Large Language Models (LLMs) hold immense potential for a variety of writing and planning tasks. Pre-trained on large text corpora, they can achieve remarkable performance on natural language tasks (Chen & Yih, 2020). However, the widespread use of LLMs has largely been limited to manually typing prompts into tools like ChatGPT, with limitations on scalability and reliability of output (Zamfirescu-Pereita et al., 2023).

Base models, such as BERT, can be fine-tuned with task-specific datasets to recognize, summarize, translate, and generate natural language text. (Devlin et al., 2018). However, the current prerequisite of a technical background significantly limits this kind of integration of LLMs into non-technical users' personal workflow (Kinnula et al., 2021).

This barrier to usability highlights the current gap between the effectiveness of AI tools and the growing pains in the usability of AI tools within the everyday workflow (Amershi et al., 2019).

In response to this challenge, this study focuses on the development of a user-friendly interface in the context of improving curriculum planning at Metropolia University of Applied Sciences. The objective is to incorporate an LLM-based tool as a part of the natural workflow of curriculum planning.

Its usability is reviewed using the cognitive walkthrough approach and a series of user tests (Polson et al., 1992). The results of these user tests will be discussed to assess the user experience and gather feedback for further refinement of this LLM-based tool as a natural part of the education workflow.

2 Text processing using NLP

Machine learning algorithms can be trained to process large volumes of data, recognize patterns and categorize information, replacing the need for human involvement in such repetitive tasks (Amershi et al., 2019).

One form of machine learning is Natural Language Processing (NLP). NLP-based solutions have gained significant traction in the field of education, as working with curriculums and study materials involves processing large amounts of text only understandable by domain experts. Data analysis, based on NLP methods, can provide a more accurate understanding and analysis, and therefore further processing, of such a large amount of data and information. (Chowdhury, 2023.)

2.1 Personalized approaches with NLP

As the field of NLP continues to evolve, there is also a growing emphasis on personalization and adaptation techniques, aiming to process natural language in a more flexible and user-specific manner (Lucie, 2020). This shift towards personalized NLP is seen as a means to address the limitations of traditional NLP tools and to provide more accessible and more effective tools for a diverse range of users. (Flek, 2020.)

The workplace environment is increasingly inundated with information, leading to entire fields of science observing information overload and its impact on employees (Parasuraman, 2011). This information overload stems from various sources, including excessive information supply, multitasking, and inadequate workplace information infrastructure (Kirsch, 2000). Such literature underscores

the need for interface designs and optimized tools that effectively manage cognitive load and information overload.

Many domain-specific tasks that involve simple categorization or data processing, for instance, could be significantly sped up through AI integration. In the context of education, this flexibility would allow study planners and curriculum coordinators to focus on higher-order tasks, such as determining the learning objectives and outcomes, rather than on the repetitive task of writing similar content.

The design and usability of web interfaces play a significant role in managing cognitive load. Insufficient contrast and task difficulty can increase cognitive load, affecting website usability (Sonderegger & Sauer, 2010). User tests for web interfaces have found selective filtering of information to be an effective way of mitigating information overload (Savolainen, 2007).

Therefore, optimizing web interface design based on the userbase's professional and personal background, and their specific cognitive load is crucial to mitigate cognitive strain and enhance user experience. Literature suggests that adaptive interfaces based on the needs of users can help reduce cognitive strain. (López-Jaquero et al., 2005.)

2.2 Modern large language models

Large language models (LLMs), like the recent GPT-4 (Achiam et al., 2023), Llama 2 (Touvron, et al. 2023) and Falcon (Almazrouei et al., 2023) have been in the spotlight in the last few years, due to their remarkable capabilities in comprehending and producing natural language content of different types. The models can understand natural language inputs and produce coherent, even intelligent, text. (Brown et al., 2020.)

LLMs are pre-trained on vast text corpora, often comprising billions of tokens, enabling them to capture a broad and diverse range of linguistic patterns and knowledge. This extensive pre-training on large-scale datasets allows LLMs to acquire a deep understanding of language, making them capable of attaining remarkable performance in different natural language tasks. (Wang et al., 2018).

When pretrained on different tasks, LLM models have demonstrated remarkable proficiency across several NLP tasks, such as machine translation (Kocmi & Federmann, 2023), question answering (Chen & Yih, 2020) and text summarization (Zhang et al., 2023).

LLM-based tools, based on PaLM (Chowdhery et al., 2022) can be customized to match the needs of curriculum writers. For one instance, these models could be fine-tuned on a vast array of educational content (See Latif et al., 2022.)

Current research has demonstrated the remarkable results of LLMs in various NLP tasks with few-shot (or even one-shot) learning, which involves inference based on a few (or just one) demonstration examples (Hegselmann, 2023). In recent years, few-shot learning has attracted substantial attention (Mosbach, 2023.)

It involves training models to make accurate predictions with only a few labelled examples. This approach is particularly useful in scenarios where labelled data is scarce or when adapting to new tasks, such as in prototyping. (Winata, 2021)

However, this prompt designing can be a laborious process. (Zamfirescu-Pereita et al., 2023) One solution to this is providing prompt templates for users to utilize, and to automatically fill in these prompts with contextual data from the user's interface environment (Cao et al., 2023). One such solution will be looked at with the tool presented in this final year project, in the context of curriculum planning.

3 Addressing the needs of curriculum development

Curriculum development is a multifaceted process. The stages of curriculum development include curriculum goal analysis, syllabus design, creating study materials, teaching students, and evaluating study outcomes. (Richards, 2001.)

Curriculum developers are responsible for identifying and incorporating these necessary goals and competencies into the curriculum to equip students for the challenges of the contemporary world. (Schwab, 1973.)

Many curriculum coordinators also act as teachers. The involvement of teachers in the curriculum development design process is crucial, as they possess first-hand knowledge of classroom dynamics and the practicality of materials. (Renfors, 2021.)

This demanding task of curriculum development requires a team of curriculum planners, or curriculum coordinators. They must know the competencies expected of students upon graduation, and the curriculum's role in helping the students attain these objectives. They are tasked with defining the expected abilities that students should possess upon transitioning to the workforce. They must guide the graduates in navigating the workforce and the surrounding world. (Alsagoff & Low, 2007.)

The development of a curriculum involves a series of complex iterative steps. Designing a curriculum requires careful consideration of various factors to ensure its effectiveness and relevance. The development of a more holistic, sustainable, and adaptable approach to curriculum design is an ongoing challenge. (Vreuls et al. 2022.)

Yet at the same time, curriculums must be updated at an ever-faster rate, with higher degree of automation, as the skills required by graduates are likely to change at a faster pace compared to previous years (Walker, 2012). The needs

of graduating students should be considered to ensure that the curriculum aligns with the competencies and skills required by employers (Pereira et al., 2020).

It must also match the current dynamic changes within the culture. The multifaceted nature of curriculum development is evident in the diverse perspectives and approaches to curriculum design, as well as the ongoing efforts to match the changes in culture and society (Green & Whitsed, 2013). This is crucial in the learning environment at every stage, as the gap on the skills and knowledge between curriculum and work life is perceptible by both students and employers (Aryanti & Adhariani, 2020).

Despite all these shared challenges, it is typical to encounter slightly varied terminologies and lexicons when a university (or any educational institution) describes its curriculum. This is because each university and university of applied sciences follows its own practices and processes in developing its curriculum and designing curriculum descriptions. (Khan et al., 2019.)

Khan et al. (2019) further highlight the importance of standards and standardized terminology in ensuring the quality and viability of curriculums. Standards for outcomes and goals are ill-defined and challenging to compare. The units of the curriculum may be programs or courses. They may comprise different kinds of units or modules and be delivered through completely different platforms, lessons or classes.

Controlling the use of a specific framework or specific vocabulary is nearly unmanageable, and there is no formal way that dictates which terms or goals are to be used. There is a need for increased alignment and a shared vision in education programs in sustainability, highlighting the need of a cohesive template to integrate shared sustainability principles into a higher education curriculum (Fraser & Bosanquet, 2006).

The UN sustainability goals offer one approach to standardized goals within higher education curriculum development (Fishman & Krajcik, 2003). The United Nations Sustainable Development Goals (SDGs) are a set of 17 global targets aimed at addressing social, economic, and environmental challenges to achieve a better and more sustainable future by 2030 (UN, 2015). Figure 1 shows the 17 different categories that make up the UN sustainability objectives.

Figure 1. The 17 UN Sustainable Development Goals (UN, 2015)

Integration of sustainable development principles into higher education curriculum is essential for preparing students to address environmental and societal challenges. By aligning the curriculum with the UN SDGs, higher education institutions can have a crucial function in creating future leaders who can address environmental and societal issues (Franco et al., 2019). Research has also pointed out how both LLMs and smaller NLP models are able to identify UN SDGs within school curriculums, making them a great standard (Kharlashkin et al., 2024).

The realisation of such progressive objectives included in official curriculum texts is far from guaranteed, even with human writers, much less machine ones. As Metropolia UAS is dedicated to including sustainability in all of its degree

programs by 2030, there is a lot of internal motivation for finding solutions to analysing sustainability within curriculums (Metropolia, 2021). At EU-level, different degrees also have certain standardized expectations for what must be included within each curriculum (EU, 2018).

Despite these guidelines, the lack of structure around the curriculum design process continues to be a problem. The notion of a curriculum is broad and dynamic, reflecting the complexity of its design process, underscoring the significance of curriculum coordinators in driving educational reforms. Curriculum coordinators play a crucial role in educational institutions, particularly in the integration of new innovations and trends into the curriculum policies.

These traditional methods of curriculum design can be laborious and time-consuming, involving numerous different tasks, including needs analysis, goal setting, syllabus design, materials development and material adaptation. Educators have historically been responsible for developing and revising learning materials as curriculums develop a process that can be repetitive and inefficient. (Voogt et al., 2019)

Could curriculum design and analysis be automated using machine learning? At least it can be sped up using many of the already existing data analysis and NLP tools, reducing the workload involved in the design process (Hamam & Loucif, 2009; Teixeira, 2020.)

LLMs' remarkable ability to understand and generate text-based content streamlines various NLP tasks (Brown et al., 2020). As earlier, NLP methods can be used to conduct content analysis of curriculum materials, identifying the general structure, recommendations, and expected learning outcomes. LLMs could potentially offer degree program coordinators an even more powerful means to identify and interpret the goals embedded within curriculum texts. (Teixeira, 2020.)

But these complex dynamics within teams of curriculum coordinators underscore the challenges and intricacies involved in realizing any form of automated approach to curriculum writing. To identify the goals within a university curriculum text corpus, it is essential to consider interdisciplinary approaches, and alignment with broader educational objectives. These tools must reduce cognitive workload and assist the curriculum planners, not add further complexity to the workflow. (Voogt et al., 2019.)

4 LLM curriculum tool and usability tests

4.1 A user-friendly application

While ChatGPT has found popular use, most NLP engines require knowledge and expertise in machine learning, natural language processing, and programming, their uptake is relatively low among laypeople (Kinnula et al., 2021).

Machine learning skills, while common among computer scientists and data analysts, are not typically found among educators. This mismatch between the required skills and those possessed by educators creates a significant gap, preventing the full integration of LLMs into educational settings. (Lindner et al., 2019.)

As different NLP tools become more and more widely used in education, their usability becomes a critical question. The technical performance of NLP tools has traditionally been evaluated numerically using objective and constraint functions and various benchmarking platforms (Wang et al., 2018). But usability of NLP tools is a question that has been approached less.

The implementation of AI should aim to ensure equitable access to these technologies. The lack of research into usability and accessibility that a wider AI

integration demand is a valid concern, especially as AI is increasingly integrated into an increasing number of work tasks. For both employers and industry, guaranteeing ease of adaptation is critical. (Zheng et al., 2015.)

By leveraging LLM-based tools, curriculum writers could generate, modify, and customize content more efficiently, thereby reducing redundancy and increasing productivity. (cf. Teixeira, 2020.)

Therefore, the design and implementation of user-friendly interfaces, as well as the evaluation of usability, will continue to be central to the development of these tools. (Kocielnik et al., 2019.)

As part of a larger remodelling of Metropolia University of Applied Sciences curriculum structure, the organization wanted to harness the potential of LLMs in curriculum design. For this a more user-friendly interface was wanted.

The aim of this final year project was to create an easy-to-use interface that combines backend LLM prompt templates and regular expressions to standardize the generated text into a data table of numbers and descriptions, with a front-end that is easy to use and reduces cognitive workload.

The application is divided into a frontend and a backend. The React backend is an academic data exploration webpage, designed to provide analysis and insights into the curriculum of the chosen domain. The color choices and visual aesthetics of the frontend were designed to resemble the Metropolia UAS brand, in order to make it feel less intimidating (cf. Sonderegger & Sauer, 2010).
The application is presented in in Figures 2 – 4.

As seen in Figure 2, the landing page of the application allows users to select a specific year and language (Finnish or English) to explore and analyse the courses across the last 20 years. Users can view detailed information from each course, including course name, credits, course content and objectives. Results

of the AI analysis, whether newly queried or earlier cached results, are visible directly within the application.

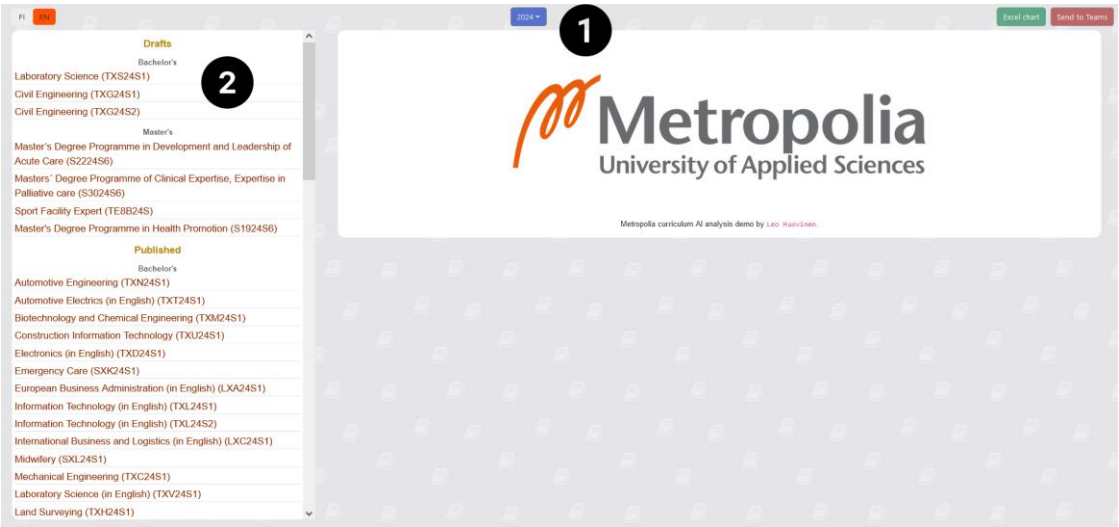

Figure 2. The landing page of the application

Orientation-wise, the top toolbar includes a dropdown menu for the years, buttons for language selection, and a download option for exporting data. A sidebar, as seen in Figure 3, contains courses of the current year, and the main page contains the curriculum to be analysed.

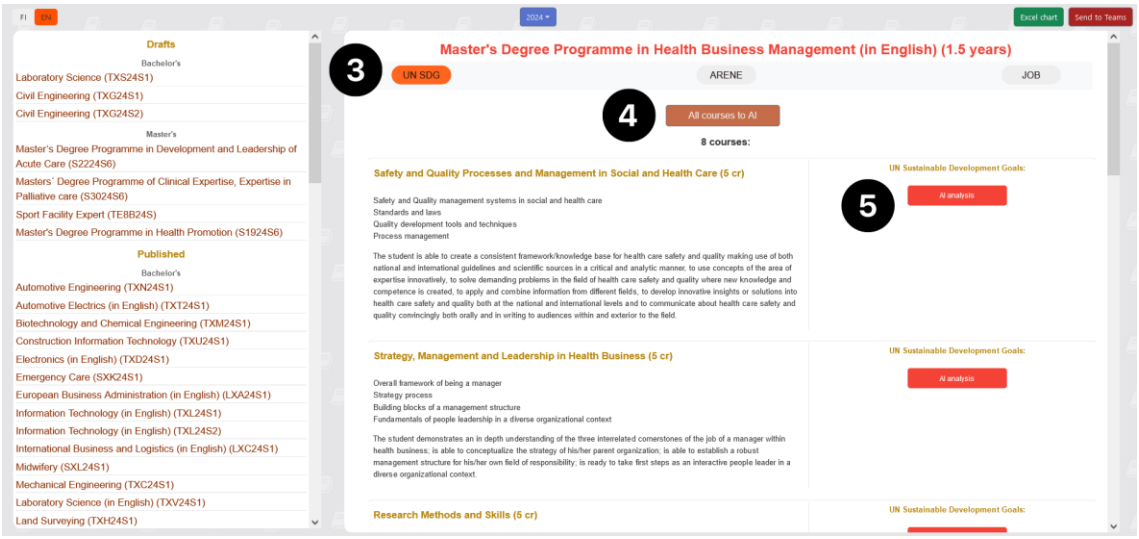

Figure 3. The curriculum analysis page of the application

As seen in Figure 4, users can further visualize the results of the LLM analysis, by using charts. Additionally, the application features loading animations and smooth scroll animations for a comprehensive user experience.

Figure 4. Results display on the curriculum analysis page of the application

The steps of using the application are represented by the following numbers:

1. Upper bar / tooltip. Always visible. The user chooses the tool language using buttons and the year of the curriculum through a dropdown menu.
2. Sidebar. Always visible. Holds all the degrees from a chosen year that the user can choose from. The user chooses their respective course curriculum domain and degrees they want to process, from a scrollable sidebar.
3. A toolbar for the goal buttons. The tool can be switched to analyze the curriculums using either UN SDG goals, the Finnish university of applied sciences goals (ARENE goals) or job market goals.
4. A button to launch the AI tool, at the center of the page. It queries all the courses of the current degree program to the Vertex backend, saving the results into a cache and into the server's MongoDB database.
5. A button to launch the AI tool for each individual course. Given the unpredictable nature of LLM's, errors and missing output may occur. Here the user can resend the query with different stochastic parameters.
6. Two buttons to upload and export the results. Initially grayed out, once a course has been selected, the user can upload the goals to the organization's backend or export the analyzed goals into an Excel spreadsheet.

Behind the interface, there is an extensive backend. Using a Flask backend (Pallets, 2023) the users can analyse the curriculum directly within the application for associated goals, such as UN SDGs, ARENE goals (Arene, 2022) and Finnish job market goals.

The backend utilizes a large language model for text generation to analyse the curriculum descriptions provided by the users. It generates a response in the form of a JSON file containing the top 5 most relevant UN SDGs based on a relevancy

score assigned to each goal. Additionally, the backend can provide a description of how the curriculum aligns with specific UN SDGs by matching quotes from the curriculum to relevant keywords.

The backend is structured to handle requests in both Finnish and English, allowing users to interact with the system based on their language preference. This is due to PaLM 2's multilingual capabilities. (Anil, 2023) The generated output is cached in a MongoDB database, providing a structured way to organize and access the analysed information for future reference, before uploading it to the organization's backend. The curriculum data itself is the latest curriculum data from the organization's Peppi database.

For the LLM part of the backend, Google Vertex AI platform's multilingual PaLM 2 model was chosen. Google Vertex AI is an integrated platform for developing, deploying, and maintaining machine learning models. This environment is designed to streamline the ML workflow. (Google, 2024) Several alternatives to Google Vertex AI, and its provided models exist in the market, each with their strengths and weaknesses. Microsoft's Azure Machine Learning is one such commercial alternative (Microsoft, 2024), open-source models such as Llama are another (Touvron, et al. 2023).

PaLM 2 is an improved language model over its predecessor, PaLM, with enhanced multilingual and reasoning capabilities. It is more compute-efficient, exhibits better quality on downstream tasks across different model sizes, and offers faster and more efficient inference. It delivers groundbreaking results, outperforming traditional state-of-the-art models on a number of multi-step reasoning language tasks. (Anil, 2023.)

Prompt templates used in the tool are structured using a one-shot strategy. Using Flask, the prompt is automatically formulated to ask the model to score the top 5 most relevant UN SDGs based on a relevancy score within a given curriculum description. This automatic prompt generation guides the model to generate

responses to automatically include appropriate course information from the browsed curriculum, requiring no prompt engineering on the user's end. (cf. Cao et al., 2023.)

The prompt also instructs the model to return the information in a specific JSON format. Regex (regular expression) processing is used to extract UN SDG goal numbers from the JSON string the model generates. This allows the tool to organize and present the information in a structured manner, utilizing data visualization libraries such as Chart.js. (Chart.js, 2023.)

Usability-wise, the app features a number of interactive elements. There are dropdown menus for selecting the year, language change buttons, and task-specific buttons for analysing different types of goals associated with the courses. Users can also download an Excel chart based on the selected course and curriculum. In the near future, this tool can also send and receive data directly between the organization's internal backend services, namely the Peppi service (Peppi-Konsortio, 2023) and Metropolia's Teams group.

With so many interactive elements and a complicated motivation for the tool, how can one be sure it is usable comfortably? Cognitive walkthrough analysis was used along with user tests to assess the accessibility and usability of these interactive elements.

4.2 Cognitive walkthrough

Cognitive walkthrough is a theory-based evaluation of user interfaces, providing a systematic way to assess the steps involved in using the webpage or program without the need for test users. The cognitive walkthrough method is a valuable approach for evaluating the usability of software and online tools, particularly in terms of identifying and predicting usability problems. It provides a structured approach to understanding user interactions, identifying usability issues, and

ultimately enhancing the user experience in a way that minimizes cognitive load. (Polson et al., 1992.)

In the context of software and online tools designed to manage cognitive load newer variations on the cognitive walkthrough method have been instrumental in assessing the usability and user experience (Mahatody et al., 2010). The Cognitive Walkthrough for the Web (CWW) is one such variant, specifically tailored for assessing the effectiveness of websites in aiding users with navigation and information retrieval (Blackmon et al., 2002).

By simulating and assessing users' internal cognitive models for specific tasks or situations, the cognitive walkthrough method can help in identifying potential error situations, information overload issues and overly complex interfaces. Furthermore, it can aid in iteratively designing and refining software and online tools to align with new users. (Kirsch, 2000.)

Cognitive workload refers to the mental exertion or focus dedicated to completing a task. Reducing cognitive workload is crucial for enhancing performance and decision-making. We have brought up earlier literature about how personalized NLP solutions could help reduce this information workload. (Flek, 2020) The goal is to help coordinators seek guidance from LLMs in formulating and refining program goals, with minimal cognitive strain.

The cognitive walkthrough was carried out using a local demo. In actual use, the application will be deployed onto a server, with the React frontend, along with the Flask (Pallets, 2023) and MongoDB (MongoDB Inc, 2023) cache existing as services accessible on the organization’s server for the users.

The cognitive walkthrough was performed with an imagined curriculum planner who wanted to analyse the 2024 curriculums for a bachelor’s degree in nursing at Metropolia UAS. While the imagined user does not have any background in

LLM, their familiarity with writing study programs, along with the goals being assessed, would guide their attitude and approach.

The steps of the cognitive walkthrough roughly correlate to each of the six interactive elements mentioned earlier:

1. The user will open the tool and know what the tool is for.
2. The user will be able to navigate to a degree program draft they are working on, or an older degree program they want to analyse.
3. The user will be able to identify the goals they want to analyse within the degree program.
4. The user will be able to run the LLM backend through all the courses within the degree program.
5. The user will be able to identify outliers generated by the LLM, mistakes or missing output, and run the LLM again with different parameters.
6. The user will be able to upload the results to the organization backend and export the results to an Excel spreadsheet.

For each of the above steps, a usability question was raised, to make sure the hypothetical user is able to complete each step. In Table 1, the usability problems that might arise during the hypothetical user's simulated walkthrough are raised, alongside potential improvements for these problems.

Table 1. Cognitive walkthrough, its expected results and added improvements.

Test question	Expected result	Improvements added during cognitive walkthrough
1. On the front page, will the user identify what the application has been designed for?	**Yes**	- No changes.
2. Can the user identify which curriculum the current in-progress draft is, and which version is from an earlier year?	**Yes**	- Courses categorized into drafts and published ones.
3. Will the user correctly identify the given goals to analyse?	**No**	- Additional color-coding and better button labels.
4. Will the user be able to successfully call the LLM and understand what the results are?	**Yes**	- Additional color-coding and better button labels.
5. Will the user navigate to outlier/missing AI results, and run the LLM again with different parameters?	**Yes**	- No changes.
6. After AI has returned all the goals, will the user navigate to the upload and export buttons?	**No**	- Upload button moved to the tooltip for better visibility.

Changes were made to the application according to the results of the cognitive walkthrough (cf. Mahatody et al., 2010).

4.3 User tests

As Metropolia University of Applied Sciences is going through a remodelling of its curriculum structure, user tests at this stage are crucial, as it was important to make sure the tool responds to specific needs of the staff. As discussed earlier, reducing cognitive load for a task as complex as curriculum development, especially during such moments of large structural change, is critical.

Structured test cases, combined with unstructured questions, were seen as the most effective way to gather data about the usability of this tool. This allowed exploring the curriculum planners' personal experiences, thoughts, attitudes, and perceptions.

The user tests were conducted in February 2024. The selected interviewees were curriculum coordinators representing various disciplines at Metropolia University of Applied Sciences. The tests were conducted within the office spaces at Metropolia UAS’s Myllypuro Campus, using a laptop platform the test users were asked to navigate the demo on.

Test subjects ranged from new curriculum planners to more experienced ones. They were chosen from various study fields, as the development goals of each are quite different. Given the variety of educational domains from which the user subjects were from - healthcare studies, architecture studies, therapeutic studies - it can be assumed the test subjects cover a range of possible users well. Within usability research, it has been found that five test subjects can identify 80% of usability problems, with extra participants unlikely to reveal new data. (Virzi, 1992.)

Test users were tasked to find and summarize sustainability accomplishments within the curriculums. They had a rough idea that this tool was going to be used to analyse sustainability and work life goals within curriculums and they knew what format the output would be in, but before the test they had never seen this application, nor did they know anything about the backend, nor what kind of AI was working in the backend. They also each held a lot of industry-specific knowledge about writing study programs and the goals analysed by the tool.

The tests were structured around the same task steps as the cognitive walkthrough, accompanied with free conversation and questions. The questions were chosen to roughly approximate a normal use case for a curriculum planner who has never seen this tool before. Each test user was asked to use the tool to navigate through a curriculum analysis task of their own domain of expertise. They were asked to analyse the 2024 curriculums according to the UN goals, internal goals and workplace goals and send them to the internal service, using the mock-up upload button.

Table 2. User test questions, and the collected results from all the test cases.

Test question	User responses (n = 5)	Suggestions for improvement
1. On the front page, can the user identify what the application has been designed for?	Yes - 2 **No - 3**	- Describe the use cases for the application on the front page.
2. Can the user identify which curriculum the current in-progress draft is, and which version is from an earlier year?	Yes - 2 **No - 3**	- Action guide on the front page. - More visible visual “year” and “draft” signifiers.
3. Will the user correctly identify the given goals to analyse?	**Yes - 5** No - 0	- Users themselves suggested further domain-specific goals to be added.
4. Will the user be able to successfully call the LLM and understand what the results are?	Yes - 4 No - 1	- Describe what the prompt is based on, in more detail.
5. Will the user navigate to outlier/missing AI results, and run the LLM again with different parameters?	**Yes - 5** No - 0	- No suggestions.
6. After AI has returned all the goals, will the user navigate to the upload and export buttons?	Yes - 2 **No - 3**	- Action guide on the front page. - An additional pop-up, when the AI is done.

Dyring step 1, several test users said that the front page of the application did not orient them towards the task enough. They suggested it is a problem that could be eased by additional information and guidance within the tool itself:

> *There is no clarification here, I wouldn't know what this is. It wouldn't hurt to have an action guide.*

> *If it's not an everyday tool, then you wouldn't have to remember every time where to click. Then it is a joy to use and does not burden my work so much.*

> *Computer tools are not my favorite thing to do. I would like a guaranteed clarification of what I need to do with just a glance, and then when I do that, the next glance provides clarity on what to do next.*

Some users also did not find or notice the year button, and when asked about it, they said it could be larger, or pointed out to the user within a step-by-step action guide.

For steps 3. and 4, they emphasized the importance of already having background working with the goals:

> *First, I notice from the button labels that this is going to tell me something about all these goals. That background information, of me working with these before is important, otherwise I might not have understood what these are used for.*

For steps 5 and 6, several test users urged caution with these generative text descriptions of program goals, as the generated descriptions for the goals are often very vague and general. For example, “Student master the methods of teamwork.” Test users felt many of these generated goal descriptions as “awkward”.

Many emphasized how important a manual human verification is, as part of this analysis work. As Vertex AI is a closed platform, the decisions made by the AI are not transparent or explainable. Verifying the goals found within the curriculums requires domain expertise.

> *Within our healthcare domain, for example, there is a world organization and EU-level requirements that sets minimum standards for education. There are specific training hours and certain areas of expertise that the students must meet. Then we have European-level skills.*

> *I hope we can spend the most time on industry-specific goals, as no one else can do that. I'd like the easiest available tool for these general overhead tasks, to avoid having to do this kind of general work. We can then focus on our own expertise.*

One test user believed that AI is not able to detect all the “weak signals” within the curriculum descriptions. A domain expert is able to recognize teaching related problems and opportunities in, especially ones related to learning important work skills.

On the other hand, one user suggested that curriculum planners themselves might not have enough information about the UN SDGs and their specific requirement. They expressed doubt whether they are able to assess the validity of these goals with a quick glance. They expressed trust in the assessment of the AI about these goals.

The users were also asked about their general perception of AI, and how they feel about the potential automation of the curriculum design process. This

human-AI connection element is something that cannot be understood through cognitive walkthrough.

During previous years, curriculum planners have used Microsoft Excel, where entering manual comments and analysis from the data was very laborious. The test users were asked about their experiences with filling out the curriculums before, and what feelings this evoked.

For some, creating curriculum plans every year was laborious, as it required a lot of writing across different platforms and there are many study program goals that need to be met. Most were motivated to try out a new tool.

> *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.*

Several test subjects emphasized cognitive workload, that comes from having to use many different platforms and tools.

> *There is a huge amount of information in different databases and tools in this house, but it is always difficult to find. Because I use them so rarely, and there are so many of them, it always takes a long time to find what I need.*

Most had used LLMs before in the form of conversational AI but felt that the guidance they had received at their work organization had been at a very general and abstract level and did not prepare them to utilize AI or LLMs effectively at work.

All the test users had used AI before, most notably ChatGPT. Many had tried ChatGPT to identify curriculum themes before. While these experiences with AI were positive, two common complaints rose. The goal analysis produced by

ChatGPT was in an unexpected format and difficult to standardize. It does not provide numbers in a comparable or aggregable format. Secondly, there was a need to write the prompt over and over, to get the desired result:

> *I have used AI (ChatGPT) in our teaching domain before to brainstorm about sustainable development goals. 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."*
>
> *If this had been the tool we used last fall, we would now have all the necessary goals ready in there already.*

Many program directors wanted the tool to be used for data exploration as well, as part of their teacher guidance tasks:

> *The tool could divide degree programs thematically and by domain. Shared courses across fields could be filtered out.*
>
> *It should be possible to compare degree programs in the same domain with each other, thereby enhancing internal communication within the organization. Then we would have a better idea how we're doing compared to others.*

Several were curious whether the tool could include a small text editor, to reduce the dependency on separate tools for writing and analysis. The tool could potentially be improved by integrating it with text editing, to include a feedback loop where the user could use the NLP to quickly iterate the development. Such web interface solutions already exist for NLP-based annotation tools (See Frasnelli, 2021.)

5 Discussion and conclusions

This final year project focused on the development of a user-friendly interface for an LLM-based curriculum tool. Its usability was reviewed using the cognitive walkthrough approach and a series of user tests. Several key themes emerged from this user feedback, shedding light on potential areas for improvement and enhancement of the tool.

One of the main areas of concern highlighted by the test users was the lack of clarity and guidance, when presented with the application without any background information. Test users expressed the need for additional information and instructions to orient them towards the task at hand. This highlights the difference between what was expected to happen in the cognitive walkthrough, and what really happened.

There was a clear frustration with having to use multiple platforms for a number of tasks. The cognitive workload and inefficiencies associated with having to jump back and forth between a large number of disparate tools and platforms can be mitigated by making these tools and interfaces as easy as possible to use, but this will not solve the underlying problem. This feedback underscores the importance of potential benefits of a centralized and integrated tool for curriculum design.

The generative text descriptions of program goals generated by the tool were also a point of contention among the test users.

How trustworthy are the predictions of these models? Many found the descriptions to be vague and general, emphasizing the need for human input and domain expertise in verifying and refining the generated goals. The explanations

given by LLMs to users are summaries of a potentially complicated decision process. This falls outside the scope of this study, and more research is needed.

These decisions are subject to their own biases and approximations, no matter how much data is used in training. (Hämäläinen, 2024) Biases are impossible to be known with large commercial models such as Vertex's PaLM. (Roberts et al., 2020).

Additionally, it should be noted that the responses provided to the user are simplifications of a complicated decision-making procedure. There could be various decision-making approaches, both erroneous and accurate, that lead to the same explanation on the user's end. (Kocielnik et al., 2019).

Moreover, the test users expressed varying opinions on the role of AI in curriculum design. This feedback underscores the importance of combining AI capabilities with human expertise to ensure accurate and meaningful results.

While some believed in the potential of AI to streamline the process and automate certain tasks, others expressed scepticism about the ability of AI to detect subtle nuances and signals within curriculum descriptions.

The education goals the AI gives to curriculum planners are simplified summaries of unknown sources of training data. It is essential to carefully test and proofread all use situations beforehand, to ensure the output's quality and effectiveness. However, the positive feedback highlights how much even partial automation of tasks can reduce cognitive workload with the most repetitive tasks.

Like with the tool, other studies too have emphasized the importance of tailoring the design of LLM-based tools, to consider specific populations and individual needs to enhance usability and user satisfaction. (Borsci et al., 2021) As language technology continues to play a significant role in people's lives, access to nuanced LLM tools becomes an issue of equality and equity.

While conversational, LLM-based online services have certainly grown more popular (Zhou et al., 2023), and NLP libraries and pipelines are increasingly accessible to even hobbyist programmers (Madnani & Loukina, 2020), these tools are out of reach for people who end up doing the most repetitive writing tasks. (Kinnula et al., 2021) This highlights the importance of addressing accessibility considerations in the development and deployment of NLP technology. (Sallam, 2023)

To adapt to new technologies, LLM or not, user tests, user feedback, training and resources will be needed. Proper preparation for work changes, to pre-emptively prevent cognitive strain through proper guidance and planning, will improve faculty members' commitment, and faculty adaptation of new technologies. (cf. Getchell et al., 2022)

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., ... & Penedo, G. (2023). The falcon series of open language models. arXiv preprint arXiv:2311.16867.

Alsagoff, L. and Low, E. (2007). Challenges in curriculum development. Relc Journal, 38(2), 229-246.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., … & Horvitz, E. (2019). Guidelines for human-ai interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. arXiv preprint arXiv:2305.10403.

Arene, (2022). Recommendation on the shared competences of universities of applied sciences and their application. Available at: https://www.arene.fi/wp-content/uploads/Raportit/2022/Kompetenssit/RECOMMENDATION%20ON%20THE%20SHARED%20COMPETENCES%20OF%20UNIVERSITIES%20OF%20APPLIED%20SCIENCES%20AND%20THEIR%20APPLICATION.pdf?_t=*1642539550* *(accessed February 23rd, 2024).*

Aryanti, C. and Adhariani, D. (2020). Students' perceptions and expectation gap on the skills and knowledge of accounting graduates. The Journal of Asian Finance, Economics and Business, 7(9), 649-657.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the web. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Borsci, S., Malizia, A., Schmettow, M., Velde, F. v. d., Tariverdiyeva, G., Balaji, D., … & Chamberlain, A. (2021). The chatbot usability scale: the design and pilot of a usability scale for interaction with ai-based conversational agents. Personal and Ubiquitous Computing, 26(1), 95-119.

Brown, T. B., Mann, B. F., Ryder, N. C., Subbiah, M., Kaplan, J., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Cao, J., Li, M., Wen, M., & Cheung, S. C. (2023). A study on prompt design, advantages and limitations of chatgpt for deep learning program repair. arXiv preprint arXiv:2304.08191.

Chart.js, (2023). Chart.js. Available at https://www.chartjs.org/docs/latest/ *(accessed February 23rd, 2024).*

Chen, D. and Yih, W. (2020). Open-domain question answering. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., … & Fiedel, N. (2022). Palm: scaling language modeling with pathways.

Chowdhury, G. (2003). Natural language processing. Annual Review of Information Science and Technology, 37(1), 51-89.

Chowdhury, G. (2003). Natural language processing. Annual Review of Information Science and Technology, 37(1), 51-89.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding.

EU (2018). The European Qualifications Framework. Available at https://europa.eu/europass/en/europass-digital-tools/european-qualifications-framework *(accessed February 23rd, 2024).*

Fishman, B. J. and Krajcik, J. (2003). What does it mean to create sustainable science curriculum innovations? a commentary. Science Education, 87(4), 564-573.

Flek, L. (2020). Returning the N to NLP: towards contextually personalized classification models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Franco, Isabel & Saito, Osamu & Vaughter, Philip & Whereat, J. & Kanie, N. & Takemoto, K.. (2019). Higher education for sustainable development: actioning the global goals in policy, curriculum and practice. Sustainability Science. 14. 10.1007/s11625-018-0628-4.

Fraser, S. and Bosanquet, A. (2006). The curriculum? that's just a unit outline, isn't it?. Studies in Higher Education, 31(3), 269-284.

Frasnelli, V., Bocchi, L., & Aprosio, A. P. (2021). Erase and rewind: manual correction of nlp output through a web interface. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Confer.

Google (2024) Vertex AI documentation. Available at https://cloud.google.com/vertex-ai/docs *(accessed February 23rd, 2024).*

Getchell, K., Carradini, S., Cardon, P. W., Fleischmann, C., Ma, H., Aritz, J., … & Stapp, J. (2022). Artificial intelligence in business communication: the changing landscape of research and teaching. Business and Professional Communication Quarterly, 85(1), 7-33.

Green, W., & Whitsed, C. (2013). Reflections on an alternative approach to continuing professional learning for internationalization of the curriculum across disciplines. Journal of Studies in International Education, 17(2), 148-164.

Hämäläinen, M. (2024). Eettisesti kestävä tekoäly. Vastuullinen hankeviestintä. Metropolia Ammattikorkeakoulu. 978-952-328-422-7.

Hamam, H. & Loucif, S., "Web-Based Engine for Program Curriculum Designers," in IEEE Transactions on Education, vol. 52, no. 4, pp. 563-572, Nov. 2009.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., & Sontag, D. (2023, April). Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics* (pp. 5549-5581). PMLR.

Khan, R., Spruijt, A., Mahboob, U., & Merriënboer, J. (2019). Determining 'curriculum viability' through standards and inhibitors of curriculum quality: a scoping review. BMC Medical Education, 19(1).

Kharlashkin, L., Macias, M., Huovinen, L., Hämäläinen, M., (2024). Predicting Sustainable Development Goals Using Course Descriptions - from LLMs to Conventional Foundation Models. arXiv preprint arXiv:2402.16420.

Kinnula, M., Iivari, N., Sharma, S., Eden, G., Turunen, M., Achuthan, K., … & Tulaskar, R. (2021). Researchers' toolbox for the future: understanding and designing accessible and inclusive artificial intelligence (aiai). Academic Mindtrek 2021.

Kirsh, D. (2000). A few thoughts on cognitive overload. Intellectica. Revue De l'Association Pour La Recherche Cognitive, 30(1), 19-51.

Kocielnik, R., Amershi, S., & Bennett, P. (2019). Will you accept an imperfect ai?. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. arXiv preprint arXiv:2302.14520.

Latif, E., & Zhai, X. (2024). Fine-tuning chatgpt for automatic scoring. Computers and Education: Artificial Intelligence, 100210.

Lindner, A., Romeike, R., Jasute, E., & Pozdniakov, S. (2019). Teachers' perspectives on artificial intelligence. In 12th International conference on informatics in schools. "Situation, evaluation and perspectives", ISSEP.

López-Jaquero, V., Montero, F., Molina, J. P., González, P., & Fernández-Caballero, A. (2005). A seamless development process of adaptive user interfaces explicitly based on usability properties. Engineering Human Computer Interaction and Interactive Systems, 289-291.

Lucie F. 2020. Returning the N to NLP: Towards Contextually Personalized Classification Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7828–7838, Online. Association for Computational Linguistics.

Madnani, N. and Loukina, A. (2020). User-centered & Robust NLP OSS: lessons learned from developing & maintaining rsmtool. Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS).

Mahatody, Thomas & Sagar, Mouldi & Kolski, Christophe. (2010). State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions. Int. J. Hum. Comput. Interaction. 26. 741-785. 10.1080/10447311003781409.

Metropolia (2021) Strategy 2021 - 2030: A bold reformer of expertise and an active builder of sustainable future. Available at https://www.metropolia.fi/en/about-us/strategy-2030/sustainable-development-and-growth *(accessed February 23rd, 2024).*

Microsoft (2024) Azure documentation. Available at https://learn.microsoft.com/en-us/azure/?product=popular *(accessed February 23rd, 2024).*

Milne-Ives, M., Cock, C. d., Lim, E., Shehadeh, M. H., Pennington, N. d., Mole, G., … & Meinert, E. (2020). The effectiveness of artificial intelligence conversational agents in health care: systematic review. Journal of Medical Internet Research, 22(10), e20346.

MongoDB Inc (2023). MongoDB documentation. Available at https://www.mongodb.com/docs/ *(accessed February 23rd, 2024).*

Mosbach, Marius & Pimentel, Tiago & Ravfogel, Shauli & Klakow, Dietrich & Elazar, Yanai. (2023). Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. 12284-12314. 10.18653/v1/2023.findings-acl.779.

Pallets (2023). Flask documentation. Available at https://flask.palletsprojects.com/en/3.0.x/ *(accessed February 23rd, 2024).*

Parasuraman, R. (2011). Neuroergonomics. Current Directions in Psychological Science, 20(3), 181-186.

Peppi-Konsortio (2023), Peppi. Available at https://www.peppi-konsortio.fi/ *(accessed February 23rd, 2024).*

Pereira, E., Vilas-Boas, M., & Rebelo, C. (2020). University curricula and employability: the stakeholders' views for a future agenda. Industry and Higher Education, 34(5), 321-329.

Renfors, S.-M. (2021). Internationalization of the Curriculum in Finnish Higher Education: Understanding Lecturers' Experiences. Journal of Studies in International Education, 25(1), 66-82.

Roberts, A., Raffel, C., & Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model?. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Sallam, M. (2023). The utility of chatgpt as an example of large language models in healthcare education, research and practice: systematic review on the future perspectives and potential limitations.

Savolainen, R. (2007). Filtering and withdrawing: strategies for coping with information overload in everyday contexts. Journal of Information Science, 33(5), 611-621.

Schwab, J. J. (1973). The practical 3: Translation into curriculum. *The school review*, *81*(4), 501-522.

Sonderegger, A. and Sauer, J. (2010). The influence of design aesthetics in usability testing: effects on user performance and perceived usability. Applied Ergonomics, 41(3), 403-410.

Teixeira, André & Guerra, Aida & Knorn, Steffi & Staffas, Kjell & Varagnolo, Damiano. (2020). Computer-aided curriculum analysis and design: existing challenges and open research directions. 1-9.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

UN (2015) - Sustainable Development Goals. Available at https://sdgs.un.org/goals *(accessed February 23rd, 2024).*

Virzi, R. A. (1992). Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough? Human Factors, 34(4), 457-468.

Voogt, J., Pieters, J. M., & Roblin, N. P. (2019). Collaborative curriculum design in teacher teams: foundations. Collaborative Curriculum Design for Sustainable Innovation and Teacher Learning, 5-18.

Vreuls, J., Koeslag-Kreunen, M., Klink, M. v. d., Nieuwenhuis, L., & Boshuizen, H. P. A. (2022). Responsive curriculum development for professional education: different teams, different tales. The Curriculum Journal, 33(4), 636-659.

Walker, M. (2012). Universities and a human development ethics: a capabilities approach to curriculum. European Journal of Education, 47(3), 448-461.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Winata, G. I., Madotto, A., Lin, Z., Liu, R., Yosinski, J., & Fung, P. (2021). Language models are few-shot multilingual learners. Proceedings of the 1st Workshop on Multilingual Representation Learning.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-21).

Zheng, K., Vydiswaran, V. G. V., Liu, Y., Wang, Y., Stubbs, A., Uzuner, Ö., … & Xu, H. (2015). Ease of adoption of clinical natural language processing software: an evaluation of five systems. Journal of Biomedical Informatics, 58, S189-S196.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2023). Benchmarking large language models for news summarization. arXiv preprint arXiv:2301.13848.

Zhou, J., Ke, P., Qiu, X., Huang, M., & Zhang, J. (2023). ChatGPT: potential, prospects, and limitations. Frontiers of Information Technology & Electronic Engineering, 1-6.