Generative AI in Curriculum Design: Empirical Insights Into Model Performance and Educational Constraints

Paulina Rutecka, Karina Cicha, Mariia Rizun, and Artur Strzelecki

***Abstract*—This study verifies the ability of large language models (LLMs) to generate a curriculum and develop syllabi for specific courses. We prompted four models to generate two sets of curricula for a bachelor's degree in Economics and Management. We also generated syllabi for the courses included in the curriculum. We chose five Polish public economics universities offering those degree programs for comparison. Four LLMs were used in this experiment: ChatGPT-3.5, ChatGPT-4, Google Bard, and Gemini. Two of them are multimodal models. The study used an iterative approach, increasing the detail of the prompt in each iteration. The results show that the more specific prompt is given to the LLM, the less accurate the results are. Moreover, the experiment shows that none of the LLMs developed a complete curriculum at a level comparable to that generated by humans. However, LLMs can significantly help create a curriculum and develop syllabi by humans, provided that there is close human–artificial intelligence (AI) collaboration. The results obtained from the AI-assisted curriculum design differ depending on the model. By analyzing the differences between the tools and the real degree programs and syllabi, we determined that multimodal models are better suited for this task than older models.**

***Index Terms*—Artificial intelligence (AI)-generated content, curriculum design, generative AI (GenAI), higher education, large language model (LLM).**

I. Introduction

THE possibilities of large language models (LLMs) as tools supporting textual work at higher education level are broadly discussed. Scientists are testing the capabilities of LLMs in supporting scientific research [1], research processes organization [2], identification of knowledge gaps [3], supporting research methods [4], and text summarizing and drawing conclusions [5].

Related research emphasizes the inadequacies and flaws of LLMs in academic work [6]. Some researchers point out that LLMs do not cope with assignments requiring critical thinking [7] or communication of scientific concepts [8], and that the use of LLMs in scientific work is unethical [9].

However, the development of LLMs is inevitable, which can already be observed through 1) developing solutions supporting text and data work by technology companies and 2) developing specialized models, the so-called Research Assistants, used in biomedical research [6], research trend identification [10], to identify studies related to a topic of interest [11], text summarization and explanation of the calculations performed in these studies [12], and generation of articles titles and abstracts [13]. Researchers also claim to use LLMs in their academic work, which does not involve a creative process but only the nonscientific content generation [14].

The use of artificial intelligence (AI) to generate nonscientific content was addressed, among others, in 2018 by a team of IBM researchers [15], who developed a model dedicated to determining learning outcomes based on the educational materials content. It provoked others [16] to verify the possibility of using paulina rutecka (GPT)-4 for the same task, to use LLMs to generate curricula for education programs and syllabi for specific courses [14], [17], and to include the use of particular teaching/learning methods in the generated curricula and syllabi [18].

Since the number of available LLMs is increasing, the question arises: to what extent can they fulfill the role of assistants in creating high-quality textual content for higher education institutions (HEIs), and which model can perform the task more efficiently?

As HEI representatives, we decided to empirically verify the assumption presented in various research works [14], [17], [18]—that LLMs can generate a complete list of topics that should be covered within syllabi and even entire study programs.

To achieve this objective, we focused on generating syllabi and curricula using LLMs. By doing so, we will be able to test and compare the textual output delivered by the LLMs and, in addition, to answer three research questions (RQs).

1) *RQ1:* Is an LLM able to propose courses that are adequate for the curricula of a particular degree program?
2) *RQ2:* Is it possible to generate an entire curriculum through an LLM?
3) *RQ3:* Which LLM provides the most similar results to human-written text?

Since the authors of this article are academics working in Management and Economics, their field of interest and the area of the experiment presented are the degree programs offered at the universities of economics in Poland.

The rest of this article is organized as follows. Section II presents the results of literature review on the application

Received 14 June 2024; revised 29 January 2025, 10 May 2025, and 28 June 2025; accepted 3 July 2025. Date of publication 8 July 2025; date of current version 30 July 2025. *(Corresponding author: Paulina Rutecka.)*

Paulina Rutecka, Mariia Rizun, and Artur Strzelecki are with the Department of Informatics, University of Economics in Katowice, 40-287 Katowice, Poland (e-mail: paulina.rutecka@uekat.pl).

Karina Cicha is with the Department of Communication Design and Analysis, University of Economics in Katowice, 40-287 Katowice, Poland.

Digital Object Identifier 10.1109/TLT.2025.3587081

TABLE I
PREVIOUS RESEARCH SUMMARY

Date	LLM activity studied	Method	Assessment	Reference
Dec 2022	Full text of the scientific article	Case study	positive	[26]
Dec 2022	Classification of answers in the survey	Comparison	positive	[27]
Dec 2022	Truth of the answer	Comparison	positive	[28]
Dec 2022	Titles of scientific articles	Expert assessment/NLG evaluation metrics	positive	[29]
Dec 2022	Scientific abstracts	Blinded human reviewers/AI and plagiarism detect	negative	[30]
Dec 2022	Full text of the scientific article	Plagiarism detector	negative	[31]
Jan 2023	Higher Order Problems in Pathology	Expert assessment	positive	[32]
Feb 2023	Selection and support of the scientific method	Case study	positive	[1]
Feb 2023	Construction project schedules	Comparison	positive	[33]
Feb 2023	Citations and References	Comparison	negative	[34]
Feb 2023	Full text of the scientific article	Expert assessment/Comparison	negative	[35]
Feb 2023	Solving student exam tasks	Comparison	positive	[36]
Mar 2023	Summary of a scientific article	Comparison	positive	[5]
Apr 2023	Bibliometric analysis	Comparison	negative	[37]
May 2023	Idea generation	Expert assessment	positive	[2]
	Literature review		negative	
	Data identification and processing		positive	
	Empirical testing		negative	

of LLMs to various tasks and methods for studying LLM performance. Section III presents the theoretical background. Section IV describes the methods used to conduct the study presented in this article. In Section V, we present the results of the study. Finally, Section VI concludes this article and analyzes its limitations, contributions, and possible directions for further research.

II. PREVIOUS RESEARCH

Since the ChatGPT was made available, the possibility of its use in academic work has been the subject of various studies. Some of them are thoughts on the possible applications of this tool [3], [4], [6], [19], [20], [21], [22], while others present the answers that the ChatGPT returned in response to the selected queries, along with their assessment [23], [24], [25]. Speculations about LLMs' possibilities, opportunities, and threats are replaced by empirical studies that verify the hypotheses put forward immediately after the launch of the ChatGPT. Publications include experiments aimed at verifying the quality of LLM responses.

Using the Web of Science database, we searched the following queries in the topic section: "chatgpt research," "chatgpt academic," "chatgpt science," "chatgpt impact education," and "chatgpt university policy." We analyzed the studies published between November 2022 and June 2023 and obtained 100 results after removing duplicates. We then discarded all the results that had a type other than Article or Review. Since the topic is still rather recent, we allowed early access publications; however, we rejected databases, such as Semantic Scholar or Google Scholar. Although they could contain interesting works, they often contain papers that have not been peer-reviewed or are from low-quality journals or publishers.

We then analyzed the abstracts to identify the empirical studies, including the method, approximate study duration, and LLM assessment for a specific activity conducted, to determine for which activities the usefulness of LLMs was tested and what the overall conclusions of these studies were (positive or negative). Table I provides a summary of the works we selected for further analysis.

One of the first studies aimed at empirically verifying the quality of texts generated by the LLM was the study of Zhai [26]. This author tested the feasibility of generating a full-text scientific article, finding that ChatGPT wrote an article that was "coherent, (partially) accurate, informative, and systematic," and it took only 2–3 h to write the article. This study aimed to carry out an empirical verification of whether a tool, such as ChatGPT, could threaten the work of a scientist. The research used a case study method.

Mellon et al. [27] conducted a study to compare the classification performance of ChatGPT-3 in coding open-ended survey responses by comparing GPT-3 classification results with human results and those obtained using a supervised machine learning algorithm [support vector machine (SVM)]. The percentage of answers correctly coded by the generative pre-trained transformer (GPT) model and the SVM algorithm was compared with the answers hand-coded by a research assistant. The study showed that LLMs are satisfactory at encoding responses.

Wenzlaff and Spaeth [28] checked the consistency of responses on financial issues generated by the ChatGPT by comparing them with human responses. Chen and Eger [29] conducted a study to determine whether LLMs can generate relevant titles for research papers using abstracts as input. Their adequacy was assessed using existing evaluation metrics designed for other natural language processing (NLP) tasks, such as machine translation or summarization and human assessment.

Gao et al. [30] investigated the ability of ChatGPT to generate abstracts for scientific articles. The abstracts were analyzed using a plagiarism detection tool, an AI detector, and blinded reviewers who received mixed abstracts—those generated by the LLM and the abstracts of the published scientific papers. Aydin and Karaarslan [31] generated the content of a scientific article, which they further assessed using antiplagiarism tools and found the text to be of low originality.

Sinha et al. [32] assessed ChatGPT's higher order reasoning ability by generating answers to 100 pathology (medicine) questions and scoring them against a predefined answer key. The responses were evaluated using qualitative assessment on a scale from 0 to 5 and the structure of the observed learning outcome taxonomy. Three expert pathologists performed the evaluation. The average ChatGPT response rate was found to be four out of five.

McDonald et al. [1] aimed to test the ability of ChatGPT to conduct scientific research. They obtained information on what statistical tests can be carried out using their dataset. The first prompt indicated their dataset, and ChatGPT suggested using correlation analysis, the $\chi 2$ test, or logistic regression. This finding also indicated that having more data could improve the study results. The second prompt provided the data with column headings and asked ChatGPT to generate the R code "to perform survival analysis and calculate risk ratios as suggested by ChatGPT." The dataset contained 100 000 rows. It was found that the code had several errors, and some data were incorrectly specified. However, the researchers were able to identify these problems and report them to the tool, which made adjustments and performed the data analysis correctly. The case study method was used in this research.

Prieto et al. [33] assessed the usefulness of ChatGPT for scheduling projects by performing a simple construction project (scope of work). Six experimental participants used ChatGPT to perform the same task, formulating prompts in their own way. Then, it was analyzed whether the elements from the baseline were included in each of the schedules proposed by the ChatGPT. In addition, several parameters were tested in the implementation of the task, such as accuracy, efficiency, clarity, coherence, reliability, relevance, consistency, scalability, and adaptability. The measurement method for each parameter was explained. Despite the overall positive assessment of the usefulness of the ChatGPT, several shortcomings were identified. It was concluded that this LLM cannot replace specialists in the area.

Day [34] analyzed the accuracy of the academic citations and references generated by the ChatGPT. In response to five questions addressed to the LLM, the ChatGPT returned 16 references. The author verified these references and concluded that none of them existed.

Das et al. [38] followed the method previously presented in [32], the authors of which also took part in the research of Das et al. In the study, 96 questions were selected, and three microbiologists rated the answers on a scale from 0 to 5. The results showed that ChatGPT can answer microbiological questions with an accuracy of approximately 80% .

Cascella et al. [5] assessed the ability of ChatGPT to understand and summarize information and draw conclusions from the text in the Background, Methods, and Results sections. Their observations show that ChatGPT can effectively summarize information and suggest further actions, including literature exploration and the generation of new research hypotheses. The researchers compared the results generated by the LLMs with the original abstract conclusions of five medical research papers.

Ariyaratne et al. [35] compared articles written by ChatGPT with the actual scientific papers published in radiology journals. These articles were assessed by two independent radiologists (expert assessment method). They found that four out of five papers were inaccurate and contained fictitious references. The quality of each section in these publications was also assessed, with the Introduction section being rated the highest. Although it was found that the articles written were consistent and seemed professional, the information returned by the tool was false. Farhat et al. [37] verified the suitability of ChatGPT for bibliometric analyses by comparing the LLM results with the real bibliometric works. It was found that the LLM provided inaccurate information about authors and countries and qualified significantly fewer publications for the study (only 19%) than people analyzing the same issue.

Dowling and Lucey [2] tested the capabilities of this ChatGPT in terms of its usefulness and adequacy in the task of testing and comparing the generated output for four stages of the research process: idea generation, literature review, data identification and processing, and empirical testing. The results generated by the tool were evaluated by experienced academic authors and reviewers (expert evaluation method). Three versions were prepared and evaluated. The first was based solely on the public data possessed by the ChatGPT, the second included abstracts of thematically related studies that expanded the expert knowledge base of the tool, and the third included the specialist knowledge of the researcher who suggested possible improvements. The idea generation and data processing stages were highly rated in this study, indicating that ChatGPT may be helpful in this context as an e-Research Assistant [2].

The current studies were limited to ChatGPT based on GPT-3. The paid version of ChatGPT using GPT-4 was made available in March 2023 [39]. GPT-4 was identified as a better language algorithm with more test parameters [2]. Google Bard was made available on 21 March 2023, but only to users invited to tests. In Poland, Google Bard was publicly available on 13 July 2023. Gemini was presented on 6 December 2023 and has been available since February 2024. Gemini is the successor to LaMDA and PaLM 2. It not only is a language model but also has been enriched with the ability to understand mathematics and physics, as well as the code of popular programming languages. Both GPT-4 and Gemini are described as multimodal models, i.e., having the ability to understand and generate data other than text [40].

In terms of the quality of the information provided, GPT-3.5 and GPT-4 were compared, among others, as part of the study in which both were subjected to a test of knowledge about gastric cancer, the aim of which was to determine whether these LLMs can help to popularize knowledge about this disease and, as a result, support in consulting and detecting the disease [41]. Another study of these LLMs is the comparison of the quality of information on medicines [42]. In both cases, GPT-4 achieved better results than did GPT-3.5. When the study was conducted, there were only two publications in the Web of Science database about Google Bard, and none related to the comparison with GPT models.

Recent studies from mid-2023 have proposed very bold attempts to use LLMs, among others, to assess human behavior as part of mental health care [43], or to assess their usefulness in surgical breast reconstruction [44]. These studies are based on

the assessment of the dialog led by the authors. Considerations are also continued regarding the possibility of LLMs taking exams in orthopedics and traumatology [45], otolaryngology [46], or management [36]. In the latter [36], ChatGPT and the responses of top-performing students were compared, and it was found that ChatGPT scores are comparable to those of high performers.

The analyzed articles covered topics, such as scientific texts written by LLMs, literature reviews, student exam preparation, and course schedule planning. However, none of the included studies is dedicated to LLM assistance in developing curricula or course syllabi for HEIs.

III. Theoretical Background

Creating a high-quality study program and syllabi is a challenging and complex task. Several factors influence this process, such as the requirements of strategic documents [47], the priorities of educational programs [48], the quality of management principles [48], and the necessity of incorporating best practices and research findings [49]. A syllabus constitutes a form of contract between an HEI and a student [49]. Therefore, educational programs should be designed coherently and tailored to market needs [48], [50], creating an attractive graduate profile for employers with relevant graduate attributes [50], [51]. To achieve this goal, programs should be constantly updated [50], [51], [52], the role of new technologies should be considered, the latest research findings should be incorporated into the curriculum, and students' transversal skills should be developed [51].

The program should be ambitious, allowing students to achieve outcomes corresponding to the appropriate qualification level for a given country, according to the European Qualifications Framework [53]. In Europe, study programs should be sufficiently aligned between universities to ensure student mobility, as guaranteed by the Bologna Declaration [48], [54], through the realization of similar topics, the achievement of specific learning outcomes, and the ability to earn European Credit Transfer System credits. However, as Biggs and Tang [54] note, excessive reliance on benchmarking can negatively impact the development of the individual character of universities and the quality of education. Therefore, it is important to maintain a balance between the individual character of a university and the possibility of credit transfers [54].

In course syllabi, all the information, such as schedules for covering topics, course objectives, descriptions of assessment methods [49], information on aligned teaching strategies [54], [55], and descriptions of instructor competencies [49], [56], should be presented transparently. It is also essential to implement constructive alignment [57], [58], understood as a pedagogical practice in which teaching, learning, and assessment activities are directly linked to the intended learning outcomes achieved by students [54], [55]. The creation of syllabi and study programs relies on the expertise of teams of experienced teachers who are responsible for meeting all the criteria of these materials' development [50], [55]. Moreover, scientists emphasize the necessity of a flexible approach to building curriculum, which includes the continuous updating of content, focusing on students and their needs, creating programs that respond to economic needs, and developing competencies instead of merely accumulating theoretical knowledge [51], [61], [62].

The approach to syllabus creation has recently shifted from a content-focused approach to a learning-focused approach [59]. Some universities, such as UMass Amherst, provide collections of best practices and guidelines to build course descriptions accordingly [60]. Among the best practices for creating course descriptions is balancing the breadth and depth of the subject matter covered in classes. As Biggs and Tang [54] note, an overly broad scope will lead to superficial learning, while focusing on depth allows for the discussion of fewer topics and deep learning within a narrow scope.

In Poland, universities' functioning is regulated by the Law on Higher Education and Science [63]. The Minister of Science and Higher Education may issue detailed regulations. None of the existing documents specify the detailed elements of syllabi beyond the learning outcomes.

Operating independently of the Ministry since 1 January 2002, the Polish Accreditation Committee (PKA) has been responsible for improving the quality of education. PKA's tasks include the evaluation of universities, as well as the development and implementation of the quality management system, which aims to enhance the effectiveness of actions fulfilling PKA's mission and to ensure that statutory tasks are performed with guaranteed consistency in quality characteristics. In 2022, the PKA issued an interpretation [64] regarding the creation of syllabi, stating that they should be considered integral documents for the study program. However, the scope of these descriptions should be determined autonomously by the Senate of each university, in accordance with the principles of the Bologna system [52].

The importance of syllabus design, as well as the mentioned possibilities of using Chat-GPT in HEIs, allows the authors to claim that the research on LLMs currently has a gap that can be filled. Therefore, the authors consider this study's objective, set before (to test and compare the textual suggestions for curricula and syllabi), to be justified.

IV. Method

To meet the previously defined objective of this article, the authors used the quasi-experimental method to compare the results of four LLMs: GPT-3.5, GPT-4, Google Bard, and Gemini, in generating curricula and course descriptions (also called syllabi). The curricula were generated for two bachelor's degree programs. In comparison, syllabi were generated for eight courses (four for each of the degree programs), which are carried out at economic universities within specific degree programs in accordance with the applicable curricula. The curricula were generated by three independent academic lecturers, which is consistent with the procedure of previously conducted research in the same field [32]. Initially, the prompts were prepared by these three experts, and later, each person prepared a set of curricula with the use of one tool based on the same prompts. Each lecturer checked the generated curricula in terms of quality

and validity. After receiving the generated content of all curricula and syllabi, two more experts, who are academic lecturers and specialists in management and economics, checked their accuracy. The dataset for this study is available online.[1]

The method used in our study aligns with the approach adopted by Ehara [65], where the calculation of the coverage rate was employed. This rate is understood as the percentage of ChatGPT-generated course concepts that matched the actual concepts from the educational platform. A similar comparison method is found in a study that compares curriculum objectives offered by various universities with qualification requirements in job postings [66]. Both approaches utilized a binary determination of coverage (yes or no) and calculated the percentage coverage rate.

This study captures the possibilities of LLMs at a specific point in time when the study was conducted, i.e., in August 2023. The analysis of the capabilities of Gemini was additionally carried out on 5 March 2024. To enable the reproduction of this study, we indicate the parameters of each of the LLMs used in the study:

1) gpt-3.5-turbo, with 4096 tokens (the name GPT-3.5 is used in this article) with default settings: temperature: 0.7, max_tokens: 2048 tokens, top_p sampling: 1, frequency_penalty: 0, presence_penalty: 0;
2) gpt-4 with 8192 tokens (the name GPT-4 is used in this article) with default settings: temperature: 0.7, max_tokens: 2048 tokens, top_p sampling: 1, frequency_penalty: 0, presence_penalty: 0;
3) PaLM 2, with 1024 tokens (the name Bard is used in this article) with default settings: temperature: 0.7, max_tokens: 1024 tokens, top_p sampling: 0, frequency_penalty: 0, presence_penalty: 0; repetition_penalty: 1; no_reapet_n_grams_size: 2;
4) gemini-pro1.0, with 2048 tokens (the name Gemini is used in this article) with default settings: temperature: 0.7, max_tokens: 2048 tokens, top_p sampling: 0,95, frequency_penalty: 0, presence_penalty: 0; repetition_penalty: 1; no_reapet_n_grams_size: 2.

A. Generating Curricula

To generate the curricula, we used an iterative approach based on the work of Dowling and Lucey [2], in which three versions of output data from LLMs were generated. Since the LLMs do not create new content, unlike proposals for scientific articles, there was no need to involve reviewers in the analysis process. The curricula generated by the LLMs were compared with the existing ones of two bachelor's degree programs in Economics and Management at five public economic universities in Poland: Cracow University of Economics (UEKR), the University of Economics in Katowice (UEKAT), Poznan University of Economics and Business (UEPN), SGH Warsaw School of Economics, and Wroclaw University of Economics and Business (UEWR). These degree programs were selected for analysis because they are offered at each university. As shown in the summary of the previous research, comparing the results generated by LLMs to human work is the most common approach for researching the usefulness of LLMs. It should be mentioned that the research method did not include verification of whether the curricula generated by LLMs were consistent with the principles of constructive alignment and curriculum coherence. The generated content was only compared with the existing content. In our methodological approach, we adopted the fundamental assumption that the curricula prepared by economic universities in Poland are consistent with these theoretical principles.

Each LLM was asked to generate a curriculum in three iterations. The primary instruction was to define the role of generative AI (GenAI) and how the process of generating curricula should be followed. The introductory prompt was "You are an experienced curriculum developer with expertise in designing comprehensive and balanced degree programs for universities. Your goal is to create a detailed curriculum that ensures students gain both theoretical knowledge and practical skills. Approach the task by first outlining the structure of the program over six semesters, then allocate specific courses, ensuring a balance between lectures and practical classes. Include foundational courses early in the program and more specialized and elective courses in the later semesters. Consider the overall workload and ensure that it is manageable and logically progresses in complexity." The first version of the answer was entirely a chatbot suggestion. In the second iteration, the prompt was amended with a request to add physical education classes and foreign languages, while in the third, detailed instructions were provided to the LLMs on assigning the number of hours and the method of crediting. Proposals for degrees in Economics and Management were generated using independent browsers and accounts on the LLMs' websites to avoid the effect of AI learning based on previous tips. According to Burger et al. [4], "the history of prompts matters for the response of a transformer model and can change the model's output."

In the first iteration, the LLMs were asked to generate proposals for curricula of Economics and Management degree programs in the most general way: "Generate a curriculum for the degree program $program_name, for bachelor studies at the university of economics. Consider division into 6 semesters. Give names of the courses, the number of hours of lectures, and practical classes. Suggest only general courses." In the second iteration, a request was added to include physical education classes and two foreign languages: "Generate a curriculum for the degree program $program_name, for bachelor studies at the university of economics. Consider division into 6 semesters. Give names of the courses, the number of hours of lectures and practical classes. Suggest only general courses. Include two foreign languages and physical education."

In the third iteration, the query was the most accurate: "Generate a curriculum for the degree program of '$program_name,' for bachelor studies at the university of economics. Consider division into 6 semesters. Provide the name of the course, the number of hours of lectures and practical classes (approx. 300 hours in semesters 1, 2, 3, and about 150 hours in semesters 4, 5, 6). Include physical education and two foreign languages

[1][Online]. Available: https://zenodo.org/records/10801874

(foreign language 1 and foreign language 2) in the first two semesters. Also, add 'General elective course' in semesters 5 and 6 and 'Elective course in a foreign language' in semester 5. The number of lecture hours should be 15 or 30 hours, and the number of hours of practical classes should be 30 hours.

Suggest only general courses. Present the result in a table with columns: course name, number of lecture hours, number of practical class hours, form of credit (exam or test), semester.

The curriculum should balance between theory and practical application, including foundational courses in $courses, along with physical education and language skills in the early semesters, followed by specialized courses and electives towards the end of the program. The inclusion of general elective courses in semesters 5 and 6 allows students to explore areas of interest further or complement their $fieldname education with interdisciplinary studies. Elective courses in foreign languages and other areas enhance the breadth and depth of the educational experience, preparing graduates for diverse roles in the $fieldname field."

The following expressions were used as variables depending on the direction: $program_name (management/economics), $fieldname (management/economics), and $courses (management, strategy, and organization/production, distribution, and consumption).

None of the LLMs suggested study objectives, learning outcomes, possible educational activities, or teaching or assessment methods. While a university should include such information in the degree program description, in our study, we did not take it into consideration and, thus, did not mention it in the prompts given to the LLMs.

B. Generating Course Syllabi

When asking the selected LLMs to generate course syllabi, we used only one iteration with the same prompt: "Generate a syllabus for $course_name for the degree program $program_name." The degree program names *($program_name*) were "Economics" and "Management." We selected the courses from each of the programs *($course_name*) included in the curriculum at each of the five universities. The management programs included Statistics in management, Organizational behavior, Project management, and Human resources management courses. For Economics, we selected Econometrics, Basics of law, Economic policy, and Basics of management.

C. Assignment of Courses in the Curricula

In making the decision to determine whether the course names within the curricula are consistent, we primarily focused on the conformity of the names. Our method is analogous to the research methods adopted in studies that compare and determine the degree of coverage, such as studies comparing curriculum objectives offered by various universities with qualification requirements in job postings [66].

We considered full or partial conformity; however, since these differences were mainly at the linguistic level (e.g., well-sounding, intriguing, or marketing-friendly names), we did not differentiate points (e.g., 1 for full conformity, and 0.5 for partial conformity). In each iteration, only the 0–1 notation was used for assessing the similarity of the names of the courses, where "0" meant no similarity at all, and "1" indicated full or partial similarity. Full similarity was indicated when the name of the course suggested by an LLM was exactly the same as the one given in the university's curricula. These include, for instance, "Econometrics" (for Economics at all five universities), and "Physical culture" (for both degree programs at all universities). Partial similarity was found in more cases than full similarity. One example of such similarity could be the course "Commercial law" (at UEKR, degree program Management) and the course "Law in business" suggested by GPT-4. While the words "business" and "commerce" are not synonyms, they are related, so these courses can be considered partially (if not even fully similar). When the course was just "Law" (at UEKAT, Management) and the LLMs suggested "Law in business," we stated the lack of similarity, since the "Law" course might have a wider choice of topics than business law. In addition, we would like to mention here that the language of the curricula and the LLMs' suggestions is Polish, so not all similarities and differences in course names may be illustrated by us in this article. We present the results as percentages for normalization, since the number of courses in each university's curricula is different [e.g., 22 courses in the UEKR and 36 courses in the UEWR).

V. Results

A. Assignment of Hours

The suggestions of the analyzed LLMs for hours were random and did not follow any logic usually applied to HEIs (at least in Poland). In the first iteration, Bard indicated 30 h for each lecture and 15 h for most practical classes; in the second iteration, it stopped showing the number from semester 4. In the third iteration, most courses received 30 h of lectures and practical classes. In the first iteration, GPT-4 appointed 45 h of lectures for most of the courses, with one reaching 60 h. The practical classes also received 45 h, with random assignments. The situation was corrected only in iteration 3, where the prompt specified the hours for lectures and practical classes. The same problem was observed for GPT-3.5. Gemini did not indicate the number of hours for the courses in any iteration. Instead, it divided the courses into lectures and seminars with specific numbers (e.g., three lectures and one seminar). Therefore, we have decided not to analyze the adequacy of assigned hours any more deeply and not to compare them with the actual curricula of Polish universities.

B. Assignment of Courses in the Curricula

Table II contains the results of the analysis of the curriculum for the Economics degree program, generated by GPT-3.5, GPT-4, Bard, and Gemini, also presented in Fig. 1. For the convenience of building this table and the following ones, we abbreviated the names of the universities as follows: UEKAT—University of Economics in Katowice, UEKR—Cracow University of Economics, UEWR—Wroclaw University of Economics

TABLE II
RESULTS OF CURRICULUM ANALYSIS—ECONOMICS DEGREE PROGRAM (%)

HEI	No. of courses	GPT-3.5			GPT-4			Bard			Gemini		
		Iteration 1	Iteration 2	Iteration 3	Iteration 1	Iteration 2	Iteration 3	Iteration 1	Iteration 2	Iteration 3	Iteration 1	Iteration 2	Iteration 3
UEKAT	23	52.17	52.17	47.83	56.52	60.87	52.17	26.09	30.43	13.04	30.43	43.48	56.52
UEKR	22	59.09	54.55	50.00	50.00	54.55	59.09	18.18	31.82	18.18	45.45	59.09	68.18
UEWR	36	38.89	33.33	25.00	50.00	52.78	36.11	25.00	33.33	11.11	30.56	36.11	38.89
UEPN	32	50.00	46.88	37.50	37.50	40.63	50.00	9.38	18.75	18.75	31.25	40.63	53.13
SGHW	31	48.39	38.71	33.33	41.94	45.16	43.33	19.35	29.03	12.90	22.58	38.71	45.16

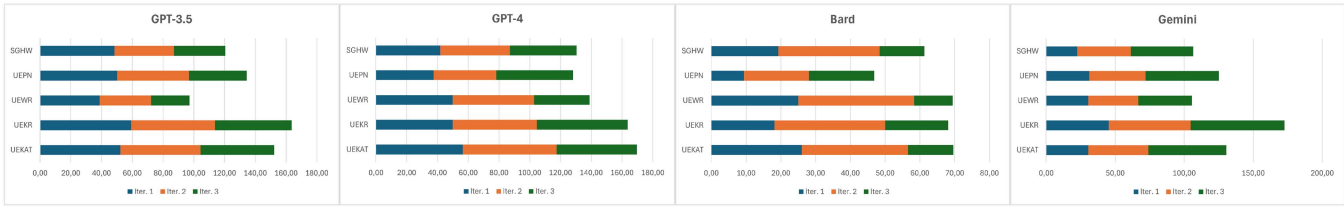

Fig. 1. Results of curriculum analysis for LLMs in iterations—Economics degree program.

TABLE III
DESCRIPTIVE STATISTICS FOR LLMS—ECONOMICS DEGREE PROGRAM

Measure	GPT-3.5	GPT-4	Bard	Gemini
Mean	44.52	48.71	21.02	42.68
Standard error	2.46	1.98	2.02	3.19
Median	47.83	50.00	18.75	40.63
Standard deviation	9.52	7.66	7.81	12.37
Minimum	25.00	36.11	9.38	22.58
Maximum	59.09	60.87	33.33	68.18

TABLE IV
DESCRIPTIVE STATISTICS FOR ITERATIONS –ECONOMICS DEGREE PROGRAM

Measure	Iteration 1	Iteration 2	Iteration 3
Mean	37.14	42.05	38.51
Standard error	3.18	2.50	3.81
Median	38.19	40.63	41.11
Standard deviation	14.24	11.18	17.05
Minimum	9.38	18.75	11.11
Maximum	59.09	60.87	68.18

and Business, UEPN—Poznan University of Economics and Business, and SGHW—SGH Warsaw School of Economics. The column "No. of courses" shows the number or courses in the actual university curricula.

In the first iteration, where the prompt was very general, both GPT models obtained better results than the Google models (Bard and Gemini)—for example, 52.17% for GPT-3.5 versus 30.43% for Gemini, for UEKAT. GPT-3.5 obtained the best result in the first iteration; in the second iteration, the best result was achieved by GPT-4, while in the third iteration, the best result was achieved by Gemini.

The courses proposed in the first iteration were often not repeated in the second and third iterations for each LLM. The answer was not supplemented with the indicated guidelines as we expected, but an entirely new proposal was generated.

Considering the average values over all iterations, GPT-4 achieved the best results, followed by GPT-3.5 and Gemini. Tables III and IV present the descriptive statistics for the courses assigned to the Economics degree program. Table IV shows the statistics for all iterations (all LLMs for all HEIs), without distinguishing between specific LLMs.

The mean and median, regardless some differences in the calculations, present the most typical values for each of the LLMs (in Table III), with the highest being for GPT-4, and present the most typical values for iterations (in Table IV), with the highest being for the second and third iterations. The lowest standard deviation, as well as the standard error, corresponds to the most consistent results: it is shown by GPT-4 among other LLMs (see Table III), and overall, in the second iteration (see Table IV).

It seems that the precision of the prompt given to the LLMs in the third iteration, and its difference from the previous prompts, resulted in the LLMs accepting a somewhat different task and thus providing results that differed greatly from the first and second iterations (which were more similar to each other).

In the second iteration, the list of courses was correctly extended to include languages and physical education classes. In the third iteration, the tools included a general elective course and an elective course in a foreign language per the given guidelines.

For GPT-3.5 and Bard, the results of the third iteration were the worst, but for the multimodal models GPT-4 and Gemini, the results of this iteration were the best. Table V presents the results of comparing the courses suggested by LLMs in three iterations for Management, also presented in Fig. 2.

TABLE V
RESULTS OF CURRICULUM ANALYSIS—MANAGEMENT DEGREE PROGRAM (%)

HEI	No. of courses	GPT-3.5			GPT-4			Bard			Gemini		
		Iteration 1	Iteration 2	Iteration 3	Iteration 1	Iteration 2	Iteration 3	Iteration 1	Iteration 2	Iteration 3	Iteration 1	Iteration 2	Iteration 3
UEKAT	25	52.00	64.00	36.00	36.00	36.00	52.00	20.00	32.00	24.00	52.00	80.00	48.00
UEKR	33	48.48	57.58	33.33	42.42	39.39	42.42	36.36	45.45	36.36	51.52	54.55	39.39
UEWR	36	50.00	58.33	33.33	44.44	41.67	38.89	27.78	36.11	30.78	55.56	58.33	30.56
UEPN	25	60.00	72.00	48.00	40.00	40.00	48.00	32.00	44.00	28.00	60.00	76.00	48.00
SGHW	30	40.00	50.00	26.67	26.67	26.67	33.33	20.00	26.67	16.67	53.33	56.67	23.33

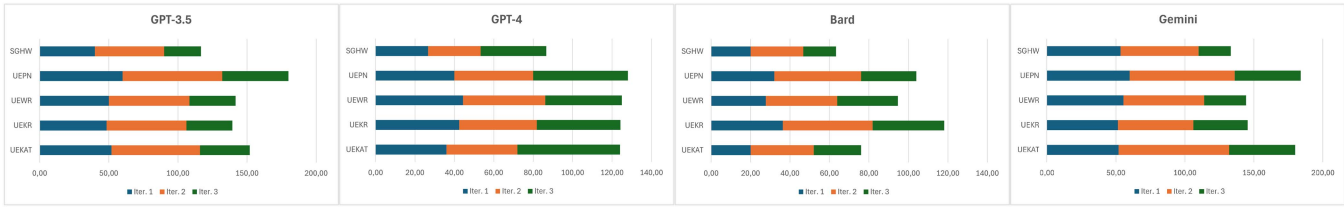

Fig. 2. Results of curriculum analysis for LLMs in iterations—Management degree program.

TABLE VI
DESCRIPTIVE STATISTICS FOR MODELS—MANAGEMENT DEGREE PROGRAM

Measure	GPT-3.5	GPT-4	Bard	Gemini
Mean	48.65	39.19	30.41	52.48
Standard error	3.30	1.78	2.17	3.76
Median	50.00	40.00	30.78	53.33
Standard deviation	12.78	6.90	8.42	14.58
Minimum	26.67	26.67	16.67	23.33
Maximum	72.00	52.00	45.45	80.00

TABLE VII
DESCRIPTIVE STATISTICS FOR ITERATIONS—MANAGEMENT DEGREE PROGRAM

Measure	Iteration 1	Iteration 2	Iteration 3
Mean	42.43	49.77	35.85
Standard error	2.78	3.50	2.18
Median	43.43	47.73	34.67
Standard deviation	12.44	15.66	9.73
Minimum	20.00	26.67	16.67
Maximum	60.00	80.00	52.00

For the Management degree program, Bard did a slightly better job than for the field of Economics. However, as with the Economics degree, this LLM never achieved 50% similarity with the actual curricula.

Gemini achieved a maximum of 80% similarity for UEKAT and 76% for UEPN in the second iteration. For GPT-4, the maximum value is 52% for UEKAT in the third iteration. For both versions of GPT, the minimum similarity percentage is 26.67%, which is better than the minimum values achieved by the Google models (16.67% for Bard and 23.33% for Gemini). In the case of generating syllabi for Management, the results of GPT-3.5 and Gemini in the second iteration were the most similar. GPT-4 performed significantly worse. After averaging, the best results were obtained by Gemini.

As in the case of Economics, the LLMs achieved the highest results in the second iteration. Descriptive statistics for the assignment of courses for the Management degree program are shown in Tables VI (for LLMs) and VII (for iterations).

While the highest result for Economics was achieved in the third iteration (see Table IV), for Management, it was achieved in the second iteration (see Table VII). The lowest standard deviation and standard error show the most consistent result for Management in the third iteration (see Table VII).

The highest standard deviation and standard error values were achieved by the Gemini model both for Economics (see Table III) and Management (see Table VI), while the lowest were achieved by the GPT-4 model, also in both cases.

This indicates that the differences between the results for different universities and across different iterations were very high for Gemini, whereas GPT-4 achieved the most consistent results.

C. *Assignment of Topics in Course Syllabi*

In this section, we compared the topics presented in the syllabi of the selected courses of the five universities with the lists of topics suggested by the LLMs. To conduct the judgment process to determine whether the thematic scope specified in the syllabi was consistent with the actual syllabi provided by HEIs, we also used a binary judgment approach. We used the 0–1 notation to assess the similarity of topics, and we primarily relied on expert assessments to determine whether the names of topics to be covered in the syllabi generated by LLMs were consistent with the actual topics in the syllabi provided by HEIs. In this case, “0” meant that the LLM suggested no such topic, and “1” meant that the LLM suggested such a topic or a partially similar

TABLE VIII
COMPARISON OF TOPICS SUGGESTED—MANAGEMENT DEGREE PROGRAM (%)

Course	GPT-3.5	GPT-4	Bard	Gemini
Statistics in management	44.78	41.79	31.34	41.79
Organizational behavior	44.93	46.38	34.78	46.38
Project management	52.94	68.63	27.45	54.90
Human resources management	52.46	59.02	50.82	54.10

TABLE IX
COMPARISON OF TOPICS SUGGESTED—ECONOMICS DEGREE PROGRAM (%)

Course	GPT-3.5	GPT-4	Bard	Gemini
Econometrics	28.99	23.19	21.74	31.88
Law	47.30	58.11	35.14	36.49
Economic policy	31.15	49.18	16.39	42.62
Basics of management	50.75	53.73	40.30	38.81

topic. Course syllabi were generated by each LLM only in one iteration.

One example of such partial similarity could be the "Organizational behavior" (Economics at UEWR) course, which includes, among other topics, "Employee motivation." The LLMs suggested the following versions of topics: GPT-3—motivation and personality; GPT-4—motivation in the workplace, key theories of motivation, motivational strategies in organizations, and motivating various groups of employees; and Bard—methods of motivating employees and motivation in the workplace. This course contains 8–26 topics (within the five universities we analyzed). This is caused, among others, by the fact that each university has its own syllabus format; for example, some write only the significant topics, while others provide a detailed list of all subtopics. We accepted the topics suggested by the LLMs as "1"—similar or partially similar. Tables VIII and IX compare the topics suggested by Bard and two versions of GPT models for four courses. The values presented are the average percentage of similarity within five universities for each of the selected courses. The Management degree program is presented in Table VIII.

GPT-4 and Gemini showed the highest similarity for three of the four courses. Only in one case, GPT-3.5 achieved better results. For Bard, the highest degree of similarity is for "Human resources management," at 50.82% . It is also the only course with a similarity above 50% for all LLMs. Judging by the fact that "Human resources management" is the only course that obtained relatively high similarity with the actual syllabi from all four LLMs, we can claim that the LLMs perceive this course as more understandable or predictable, perhaps more common for universities.

Table IX compares the topics suggested by four LLMs for four courses within the Economics degree program. In the case of this degree program, the "leading" courses are "Law" and "Basics of management." The "Basics of management" course seems to be the most "successful" course for all the LLMs. As in the case of the "Human resources management" in the Management degree program, the "Basics of management" seems to be the most predictable and unified for all four LLMs.

VI. DISCUSSION AND CONCLUSION

Since among the various possible uses of LLMs in education, generating curricula and syllabi for specific courses is one of the most frequently mentioned [14], [17], the authors' experiment aimed to verify that.

The experiment shows that none of the analyzed LLMs can develop a complete curriculum at a level comparable to that generated by humans. By analyzing the differences between the tools, we determined that multimodal models, such as the GPT-4 and Gemini, are better suited for this task than older models. However, in most cases, the results for GPT-3.5 did not differ significantly.

None of the LLMs correctly assigned the number of teaching hours to a subject based on the instructions provided. However, it can be concluded that this type of activity is not a typical task for which LLMs were created and that other algorithms, e.g., decision trees, would cope with this problem better [67]—which may be a starting point for another research. However, multimodal models, such as GPT-4 and Gemini, did not perform better than GPT-3.5, which cannot analyze data or code.

We also analyzed the differences in the results depending on the construction of the prompt. We observed that the precision of the prompts had a significant impact on the results generated by the LLMs. Simpler models (GPT-3.5 and Bard) had difficulty adapting to more complex prompts. Both GPT models performed better at generating responses to a general prompt than the Google models. GPT-3.5 achieved the best result using a very general simple prompt. More advanced models (GPT-4 and Gemini) handled precise prompts better, achieving their best results with very precise queries. However, generalizing the results of all the LLMs, the best results were obtained in the second iteration. Therefore, it can be concluded that the best results were achieved when the prompt included some specificity of the task but did not impose highly detailed criteria. Considering the described process of giving the prompts to the LLMs, this knowledge may be helpful for another try or experiment conducted using a newer model's version or in a different field of study.

Unfortunately, we found that refining the prompt to obtain better results does not always yield the expected outcome. At the level of creating curricula, depending on the iteration, LLMs added specific courses that were not included in subsequent attempts. When the queries were refined to be more detailed, the adequacy of the answers decreased for all LLMs.

Analyzing the differences between the courses, based on the average values of all iterations for generating the curriculum for the Economics and Management programs, we determined that for Economics, GPT-4 achieved the best results, followed by GPT-3.5 and Gemini. For the Management program, the best results were achieved by Gemini. The results of generating educational programs and topics covered within subjects are not satisfactory. For the most part, the LLMs' results did not match the programs developed by the university's teams of specialists. However, our results are better than those for generating course

concepts in a study by Ehara [65], where the conclusions are that the proposals generated by LLMs are not consistent with human propositions. At this point, however, it is worth noting that there are significant observable differences between the programs established by people at different universities.

According to the university guidelines, the syllabi should indicate the recommended literature. However, it is known from previous studies that LLMs are prone to hallucinations [5] and that LLMs are unable to produce learning outcomes under the assumptions made because they require abstract thinking, which they are not capable of [7]. Therefore, the results obtained in this study did not include the suggested literature or learning outcomes. It is worth noting that Google models proposed the literature at the level of creating course syllabi, but it was completely incorrect, and the suggested literature propositions do not exist. The LLMs should be familiar with a complete list of required textbooks for each course to obtain a higher level of precision in generating literature for the given subject.

Each of the analyzed LLMs made mistakes when creating curricula and syllabi for specific courses. Subsequent iterations increased the number of inaccuracies, which translated into a lower similarity percentage between the actual and generated content. Bard's results for each of the analyzed areas were the worst in terms of quality.

In summary, we conclude that none of the studied LLMs can generate syllabi and curricula completely, none is error-free, and none follows university guidelines, because this task is still too complex. Course suggestions generated by LLMs as part of curricula and content within the subjects can be considered logically justified. Therefore, we conclude that LLMs can support generating ideas or inspire academics to develop such documents, but they must be subject to human verification.

Our study results closely resemble the findings of Hoffmann et al. [68], who employed the ChatGPT to construct a psychometric scale in which the tool helped define the scale's dimensions and generate relevant items. Similarly, as in our study, the findings highlight the limitations of using AI in content generation. While GenAI effectively generated curricula and syllabi and proposed a wide range of subjects, human intervention is important for refining these subjects and confirming their validity and quality.

Some of the risks of using GenAI in curricular design are derived from GenAI tools. These include the creation of false information, bias and discrimination, intellectual property and copyright issues, privacy and security breaches, quality and reliability concerns, and compliance challenges. The tools of the GenAI can also create nonexistent literature items, which results from a tendency toward so-called hallucinations [5]. In our study, we identified the most obvious risks as generating inaccurate names of subjects, inaccurate course content, incorrect number of hours assigned for each course, illogical order of subjects in each period, risks related to plagiarism by providing unauthorized content, and overreliance on AI-generated content. In the case of using GenAI as a support for generating the course descriptions and syllabi, our recommendations stand as follows.

1) AI must be treated only as a support tool for content generation, which needs to be overseen by an expert.
2) The prompts balance is required (not too general, yet not too detailed).
3) The content requires verification in terms of consistency and potential errors.
4) Ensuring that the selected AI-based tool has the correct literature set is required.

A. Contribution of the Study

The study's contribution may be explained as follows. First, by examining the capabilities of four LLMs—ChatGPT-3.5, ChatGPT-4, Google Bard, and Gemini—in generating higher education curricula and syllabi, this study presents a novel application of GenAI in academic program design, addressing a research gap in LLMs' application to educational design.

Second, the study provides a foundation for future interdisciplinary research, underscoring the role of AI as a complementary tool rather than a replacement in academic innovation.

Third, this study allowed us to verify the assumption made by researchers in other scientific studies that LLMs can be used in academic work to generate curricula and syllabi for subjects. However, unlike the studies identified during the literature review, in our experiment, the verification was not limited to analyzing the content generated by ChatGPT. The results generated by the four LLMs were compared.

LLMs are a technology that will continue to evolve. Therefore, the study's contribution is to help indicate which of the currently publicly available LLMs can (and to what extent) generate the content necessary for creating and conducting degree programs by universities.

B. Limitations of the Study

Our study has four limitations. First, only degree programs and courses belonging to the discipline of Economics and Management were considered. However, we are aware that for other academic disciplines, such as science, technology, engineering, and mathematics (STEM) field, the results may differ. Second, there were few convergent degree programs at the analyzed universities, and within the selected programs, there were few convergent courses for which a comparison could be made. Therefore, the scale (number of courses and curricula analyzed) is small. Third, it should be emphasized that although all Polish universities operate within the framework of the guidelines adopted at the pan-European level in the Bologna Process, they have autonomy in shaping curricula and content implemented in specific courses. This, in turn, makes the input material (syllabi and curricula) used for comparison with the content generated by LLMs not homogeneous. The results of the quasi-experiment were validated by comparing them with the existing curricula and course descriptions from Polish Universities of Economics. The adequacy of the outputs generated by LLMs may also vary when compared to the actual curricula and course descriptions from universities in other countries, which operate within different educational systems.

The expert evaluation method used in the study does not contain interrater reliability measures for expert evaluations. Furthermore, we are aware of potential biases in LLM training

data. Finally, we are aware that the binary method of judgment misses to address important educational qualities like systematic course ordering and unified pedagogical structure alongside the balance of different fields.

C. Avenue for Further Research

The present study can serve as a starting point for comparing results from other disciplines or courses. It can also be a voice in the discussion on the risk of generating academic content through LLMs. Finally, the conducted study can be used as a model for conducting similar verifications regarding the assumptions about the content that LLMs can generate. In particular, it is worth repeating the study with a more extensive assessment method, not limited to binary assessment, and taking into account the broader context. The possibilities of LLMs in terms of assigning didactic hours are also worth implementing, but with the use of more developed multimodal models that have been enriched with the possibilities of logical reasoning. This research is also worth repeating using LLMs with dataset optimization coupled with a reinforcement learning approach to curriculum development.

References

[1] C. Macdonald, D. Adeloye, A. Sheikh, and I. Rudan, "Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis," *J. Glob. Health*, vol. 13, 2023, Art. no. 01003, doi: 10.7189/jogh.13.01003.

[2] M. Dowling and B. Lucey, "ChatGPT for (finance) research: The Bananarama Conjecture," *Finance Res. Lett.*, vol. 53, 2023, Art. no. 103662, doi: 10.1016/j.frl.2023.103662.

[3] K. Cheng et al., "Potential use of artificial intelligence in infectious disease: Take ChatGPT as an example," *Ann. Biomed. Eng.*, vol. 51, pp. 1130–1135, 2023, doi: 10.1007/s10439-023-03203-3.

[4] B. Burger, D. K. Kanbach, S. Kraus, M. Breier, and V. Corvello, "On the use of AI-based tools like ChatGPT to support management research," *Eur. J. Innov. Manage.*, vol. 26, pp. 233–241, 2023, doi: 10.1108/EJIM-02-2023-0156.

[5] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios," *J. Med. Syst.*, vol. 47, p. 33, 2023, doi: 10.1007/s10916-023-01925-4.

[6] L. De Angelis et al., "ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health," *Front. Public Health*, vol. 11, 2023, Art. no. 1166120, doi: 10.3389/fpubh.2023.1166120.

[7] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, "A SWOT analysis of ChatGPT: Implications for educational practice and research," *Innov. Educ. Teach. Int.*, vol. 61, no. 3, pp. 460–474, May 2024, doi: 10.1080/14703297.2023.2195846.

[8] J. Wittmann, "Science fact vs science fiction: A ChatGPT immunological review experiment gone awry," *Immunol. Lett.*, vol. 256, pp. 42–47, 2023, doi: 10.1016/j.imlet.2023.04.002.

[9] G. Tang, "Letter to editor: Academic journals should clarify the proportion of NLP-generated content in papers," *Accountability Res.*, vol. 31, no. 8, pp. 1242–1243, Dec. 2024, doi: 10.1080/08989621.2023.2180359.

[10] D. Heaven, "AI peer reviewers unleashed to ease publishing grind," *Nature*, vol. 563, pp. 609–610, 2018, doi: 10.1038/d41586-018-07245-9.

[11] M. Gusenbauer, "Audit AI search tools now, before they skew research," *Nature*, vol. 617, pp. 439–439, 2023, doi: 10.1038/d41586-023-01613-w.

[12] C. Zhang et al., "One small step for Generative AI, one giant leap for AGI: A complete survey on ChatGPT in AIGC era," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.06488

[13] E. Isaeva, "Computer-aided instruction for efficient academic writing," in *Advances in Artificial Systems for Medicine and Education V. AIMEE 2021. Lecture Notes on Data Engineering and Communications Technologies*, vol. 107, Z. Hu, S. Petoukhov, and M. He Eds., Berlin, Germany: Springer-Verlag, 2022, pp. 546–555.

[14] M. E. Emenike and B. U. Emenike, "Was this title generated by ChatGPT? Considerations for Artificial Intelligence text-generation software programs for chemists and chemistry educators," *J. Chem. Educ.*, vol. 100, no. 2, pp. 413–448, 2023, doi: 10.1021/acs.jchemed.3c00063.

[15] K. N. Tran et al., "Document chunking and learning objective generation for instruction design," in *Proc. 11th Int. Conf. Educ. Data Mining*, 2018, pp. 1–10.

[16] P. Sridhar, A. Doyle, A. Agarwal, C. Bogart, J. Savelka, and M. Sakr, "Harnessing LLMs in curricular design: Using GPT-4 to support authoring of learning objectives," 2023. [Online]. Available: http://dx.doi.org/10.48550/arXiv.2306.17459

[17] S. Ivanov and M. Soliman, "Game of algorithms: ChatGPT implications for the future of tourism education and research," *J. Tourism Futures*, vol. 9, pp. 214–221, 2023, doi: 10.1108/JTF-02-2023-0038.

[18] G. Cooper, "Examining science education in ChatGPT: An exploratory study of Generative Artificial Intelligence," *J. Sci. Educ. Technol.*, vol. 32, pp. 444–452, 2023, doi: 10.1007/s10956-023-10039-y.

[19] M. Hutson, "Could AI help you to write your next paper?," *Nature*, vol. 611, pp. 192–193, 2022, doi: 10.1038/d41586-022-03479-w.

[20] Y. K. Dwivedi et al., "Opinion Paper: 'so what if ChatGPT wrote it?' multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *Int. J. Inf. Manage.*, vol. 71, 2023, Art. no. 102642, doi: 10.1016/j.ijinfomgt.2023.102642.

[21] Y. Gendron, J. Andrew, and C. Cooper, "The perils of Artificial intelligence in academic publishing," *Crit. Perspectives Accounting*, vol. 87, 2022, Art. no. 102411, doi: 10.1016/j.cpa.2021.102411.

[22] A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, and H. Müller, "AI for life: Trends in artificial intelligence for biotechnology," *New Biotechnol.*, vol. 74, pp. 16–24, 2023, doi: 10.1016/j.nbt.2023.02.001.

[23] M. Alshater, "Exploring the role of Artificial Intelligence in enhancing academic performance: A case study of ChatGPT," Available at SSRN 4312358, 2022.

[24] D. Haluza and D. Jungwirth, "Artificial Intelligence and ten societal megatrends: An exploratory study using GPT-3," *Systems*, vol. 11, 2023, Art. no. 120, doi: 10.3390/systems11030120.

[25] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in ChatGPT: Implications in scientific writing," *Cureus*, vol. 15, 2023, Art. no. e35179, doi: 10.7759/cureus.35179.

[26] X. Zhai, "ChatGPT user experience: Implications for education," *SSRN Electron. J.*, 2022. [Online]. Available: https://ssrn.com/abstract=4312418

[27] J. Mellon, J. Bailey, R. Scott, J. Breckwoldt, M. Miori, and P. Schmedeman, "Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale," *Res. Politics*, vol. 11, no. 1, 2024, doi: 10.1177/20531680241231468.

[28] K. Wenzlaff and S. Spaeth, "Smarter than humans? Validating how OpenAI's ChatGPT model explains crowdfunding, alternative finance and community finance," *SSRN Electron J.*, 2022. [Online]. Available: https://ssrn.com/abstract=4302443

[29] Y. Chen and S. Eger, "Transformers go for the LOLs: Generating (Humourous) titles from scientific abstracts end-to-end," in *Proc. 4th Workshop Eval. Comparison NLP Syst.*, 2023, pp. 62–84.

[30] C. A. Gao et al., "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers," *npj Digit. Med.*, vol. 6, 2023, Art. no. 75, doi: 10.1038/s41746-023-00819-6.

[31] Ö. Aydın and E. Karaarslan, "OpenAI ChatGPT generated literature review: Digital twin in healthcare," *Emerg. Comput. Technol.*, vol. 2, Ö. Aydın Ed., İzmir Akademi Dernegi., 2022, pp. 22–31, doi: 10.2139/ssrn.4308687.

[32] R. K. Sinha, A. Deb Roy, N. Kumar, and H. Mondal, "Applicability of ChatGPT in assisting to solve higher order problems in pathology," *Cureus*, vol. 15, 2023, Art. no. e35237, doi: 10.7759/cureus.35237.

[33] S. A. Prieto, E. T. Mengiste, and B. G. de Soto, "Investigating the use of ChatGPT for the scheduling of construction projects," *Buildings*, vol. 13, 2023, Art. no. 857, doi: 10.3390/buildings13040857.

[34] T. Day, "A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT," *Professional Geographer*, vol. 75, no. 6, pp. 1024–1027, 2023, doi: 10.1080/00330124.2023.2190373.

[35] S. Ariyaratne, K. P. Iyengar, N. Nischal, N. C. Babu, and R. Botchu, "A comparison of ChatGPT-generated articles with human-written articles," *Skeletal Radiol.*, vol. 52, pp. 1755–1758, 2023, doi: 10.1007/s00256-023-04340-5.

[36] I. S. Chaudhry, S. A. M. Sarwary, G. A. El Refae, and H. Chabchoub, "Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT—A case study," *Cogent Educ.*, vol. 10, 2023, Art. no. 2210461, doi: 10.1080/2331186X.2023.2210461.

[37] F. Farhat, S. S. Sohail, and D. Ø. Madsen, "How trustworthy is ChatGPT? The case of bibliometric analyses," *Cogent Eng.*, vol. 10, 2023, Art. no. 2222988, doi: 10.1080/23311916.2023.2222988.

[38] D. Das et al., "Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum," *Cureus*, vol. 15, 2023, Art. no. e36034, doi: 10.7759/cureus.36034.
[39] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *npj Digit. Med.*, vol. 6, 2023, Art. no. 120, doi: 10.1038/s41746-023-00873-0.
[40] S. Pichai and D. Hassabis, "Introducing Gemini: Google's most capable AI model yet," 2023. Accessed: 9 Mar. 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/
[41] J. Zhou et al., "Exploring ChatGPT's potential for consultation, recommendations and report diagnosis: Gastric cancer and gastroscopy reports' Case," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 8, pp. 7–13, 2023, doi: 10.9781/ijimai.2023.04.007.
[42] N. He et al., "Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries," *J. Telemed. Telecare*, vol. 31, no. 2, pp. 306–308, 2025, doi: 10.1177/1357633X231181922.
[43] M. Haman, M. Školník, and T. Šubrt, "Leveraging ChatGPT for human behavior assessment: Potential implications for mental health care," *Ann. Biomed. Eng.*, vol. 51, pp. 2362–2364, 2023, doi: 10.1007/s10439-023-03269-z.
[44] Y. Xie, I. Seth, W. M. Rozen, and D. J. Hunter-Smith, "Evaluation of the artificial intelligence chatbot on breast reconstruction and its efficacy in surgical research: A case study," *Aesthetic Plast. Surg.*, vol. 47, pp. 2360–2369, 2023, doi: 10.1007/s00266-023-03443-7.
[45] R. Cuthbert and A. I. Simpson, "Artificial intelligence in orthopaedics: Can Chat generative pre-trained transformer (ChatGPT) pass section 1 of the Fellowship of the royal college of surgeons (trauma & orthopaedics) examination?," *Postgrad. Med. J.*, vol. 99, pp. 1110–1114, 2023, doi: 10.1093/postmj/qgad053.
[46] C. C. Hoch et al., "ChatGPT's quiz skills in different otolaryngology subspecialties: An analysis of 2576 single-choice and multiple-choice board certification preparation questions," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 280, pp. 4271–4278, 2023, doi: 10.1007/s00405-023-08051-4.
[47] D. Jankowska, "O wewnętrznych systemach zapewniania jakości w uczelniach polskich w kontekście specyfiki ich genezy i wdrożeń," *Pedagog Szk Wyższej*, vol. 2013, no. 13, pp. 47–61, 2013.
[48] K. Przystupa, "Jakość kształcenia w uczelni wyższej," *Autobusy Tech. Eksploat Syst. Transp.*, vol. 18, pp. 1770–1775, 2017.
[49] J. L. Wagner et al., "Best practices in syllabus design," *Amer. J. Pharmaceut. Educ.*, vol. 87, pp. 432–437, 2023, doi: 10.5688/AJPE8995.
[50] O. Vettori, *Learning & Teaching Paper# 8. Curriculum Design: Thematic Peer Group Report*. Geneva, Switzerland: Eur. Univ. Assoc., 2020.
[51] O. Hicks, "Curriculum in higher education: Confusion, complexity and currency," *HERDSA Rev. Higher Educ.*, vol. 5, pp. 5–30, 2018.
[52] S. Adam, "Towards a European higher education area curriculum development good practice guide 'strengthening higher education in Bosnia Herzegovina," (SHE III) Council of Europe European Union, 2011. [Online]. Available: https://pjp-eu.coe.int/bih-higher-education/images/curriculum_development_good_practice_guide_eng_feb2011_-_abbyy.pdf
[53] MEN GOV, "Polska Rama Kwalifikacji (PRK) i Europejska Rama Kwalifikacji (ERK)—Punkt Koordynacyjny ds," Polskiej i Europejskiej Ramy Kwalifikacji, 2011. Accessed: Jun. 12, 2024. [Online]. Available: https://prk.men.gov.pl/polska-rama-kwalifikacji-prk-i-europejska-rama-kwalifikacji-erk/
[54] J. Biggs and C. Tang, *Teaching for Quality Learning at University*. Buckingham, U.K.: Soc. Res. Higher Educ./Open Univ. Press, 1999.
[55] G. O'Neill, "Curriculum design in higher education: Theory to practice," in *Teaching and Learning*. Dublin, Ireland: Univ. College Dublin, 2015, pp. 251–253.
[56] C. Johnson, "Best practices in syllabus writing: Contents of a learner-centered syllabus," *J. Chiropractic Educ.*, vol. 20, pp. 139–144, 2006, doi: 10.7899/1042-5055-20.2.139.
[57] J. Biggs, "Enhancing teaching through constructive alignment," *High Educ.*, vol. 32, pp. 347–364, 1996, doi: 10.1007/BF00138871.
[58] C. Loughlin, S. Lygo-Baker, and Å. Lindberg-Sand, "Reclaiming constructive alignment," *Eur. J. High Educ.*, vol. 11, pp. 119–136, 2021, doi: 10.1080/21568235.2020.1816197.
[59] M. S. Palmer, L. B. Wheeler, and I. Aneece, "Does the document matter? The evolving role of syllabi in higher education," *Change: Mag. Higher Learn.*, vol. 48, no. 4, pp. 36–47, Jul. 2016, doi: 10.1080/00091383.2016.1198186.
[60] UMassAmherst Center for Teaching & Learning, "Handout video series: Six principles of an inclusive syllabus," 2021. Accessed: 13 Jun. 2024. [Online]. Available: https://www.umass.edu/ctl/sites/default/files/2021-09/HandoutVideoSeries-SixPrinciplesofanInclusiveSyllabus.pdf
[61] V. T. Shailashri, P. S. Acharya, and M. D. Pradeep, "Innovations and best practices in designing quality curricular aspects for higher education," in *Challenges Qual. Sustenance Higher Educ. Conf.*, 2019, pp. 32–38.
[62] K. R. Ramdass and K. Mokgohloa, "Curriculum design in higher education: A reflection," in *Proc. 22nd Eur. Conf. e-Learn.*, 2023, pp. 252–260.
[63] Ministry of Education and Science, "Ustawa z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce. Dz. U. 2018 poz. 1668 (ang. Act: Law on higher education and science, July 20 2018)," 2018. [Online]. Available: https://dziennikustaw.gov.pl/DU/rok/2018/pozycja/1668
[64] P. K. Akredytacyjna, "Stanowisko interpretacyjne nr 10/2022 Prezydium Polskiej Komisji Akredytacyjnej z dnia 9 czerwca 2022 r. (ang. Polish Accreditation Committee (2022) Interpretive position no. 10/2022 Presidium of the Polish Accreditation Committee)," 2022. [Online]. Available: https://pka.edu.pl/stanowiska/
[65] Y. Ehara, "Measuring similarity between manual course concepts and ChatGPT-generated course concepts," in *Proc. 16th Int. Conf. Educ. Data Mining*, M. Feng, T. Käser, and P. Talukdar, Eds., 2023, pp. 474–47, doi: 10.5281/zenodo.8115758.
[66] Y. Li, X. Wang, and D. Xin, "An inquiry into AI university curriculum and market demand," in *Proc. Comput. People Res. Conf.*, 2019, pp. 139–142.
[67] H. F. Zhou, J. W. Zhang, Y. Zhou, X. Guo, and Y. Ma, "A feature selection algorithm of decision tree based on feature weight," *Expert Syst. Appl.*, vol. 164, 2021, Art. no. 113842, doi: 10.1016/J.ESWA.2020.113842.
[68] S. Hoffmann, W. Lasarov, and Y. K. Dwivedi, "AI-empowered scale development: Testing the potential of ChatGPT," *Technol. Forecasting Soc. Change*, vol. 205, 2024, Art. no. 123488, doi: 10.1016/j.techfore.2024.123488.

Paulina Rutecka received the Ph.D. degree in management from the University of Economics, Katowice, Poland, in 2023.

She is currently an Assistant Professor with the Department of Informatics, University of Economics in Katowice. Her research interests include the impact of sustainable development communication on consumer behavior, as well as the quality of corporate communication on the Internet, including the quality of websites and analysis of search engine algorithms.

Karina Cicha received the Ph.D. degree in humanities from the University of Wrocław, Wrocław, Poland, in 2012.

She is currently an Assistant Professor with the Department of Communication Analysis and Design, University of Economics in Katowice, Katowice, Poland. Her recent scientific work has been dedicated to the satisfaction from online education and issues of communication. Her research interests include communication design, especially the field of visual communication, and media literacy among students.

Mariia Rizun received the joint master's degree from the National Mining University of Ukraine, Dnipro, Ukraine, and the University of Economics, Katowice, Poland, and the Ph.D. degree in social sciences in management from the University of Economics, in 2022.

She is currently an Assistant Professor with the Department of Informatics, University of Economics. Her current research interests include maturity models, business process modeling, knowledge management, education management, personalized education, and higher education.

Artur Strzelecki received the Ph.D. degree in management and the D.Sc. degree in management and quality sciences from the University of Economics in Katowice, Katowice, Poland, in 2013 and 2021, respectively.

He is currently an Associate Professor with the Department of Informatics, University of Economics in Katowice. His research interests include technology acceptance, information management, search engines, e-commerce, and social media.