KICK 4.0 - AI Chatting Skills in the Engineering Laboratory

Johannes Kubasch, University of Wuppertal

Johannes Kubasch is a mechanical engineer and research associate at the Chair of Technical and Engineering Education at the University of Wuppertal. As a engineer in automotive engineering, he initially worked in the automotive supply industry in the development of airbag systems before moving to the University of Wuppertal to work in the field of engineering education. In the past, he worked on the AdeLeBk.nrw project to digitize the university training of prospective teachers at vocational schools and to adapt the learning content to the requirements of later professional practice.

Dr. Dominik May, University of Wuppertal

Dr. May serves as a Professor for Technical Education and Engineering Education Research at the School of Mechanical Engineering and Safety Engineering at University of Wuppertal. His work revolves around generating both fundamental and practical knowledge that defines, informs, and enhances the education of engineers.

His primary research thrust centers around the development, implementation, practical utilization, and pedagogical value of online laboratories. These laboratories span a range of formats, including remote, virtual, and cross-reality platforms. Dr. May's scholarly pursuits extend into the sphere of online experimentation, particularly within the context of engineering and technical education. Prior to his role at the University of Wuppertal, Dr. May held the position of Assistant Professor within the Engineering Education Transformations Institute at the University of Georgia (Athens, GA, USA).

Central to Dr. May's scholarly endeavors is his commitment to formulating comprehensive educational strategies for Technical and Engineering Education. His work contributes to the establishment of an evidence-based foundation that guides the continual transformation of Technical and Engineering Education. Additionally, Dr. May is actively involved in shaping instructional concepts tailored to immerse students in international study contexts. This approach fosters intercultural collaboration, empowering students to cultivate essential competencies that transcend cultural boundaries.

Beyond his academic role, Dr. May assumes the position of President at the "International Association of Online Engineering (IAOE)," a nonprofit organization with a global mandate to advocate for the broader advancement, distribution, and practical application of Online Engineering (OE) technologies. His leadership underscores his commitment to leveraging technological innovation for societal progress. Furthermore, he serves as the Editor-in-Chief for the "International Journal of Emerging Technologies in Learning (iJET)," a role that facilitates interdisciplinary discussions among engineers, educators, and engineering education researchers. These discussions revolve around the interplay of technology, instruction, and research, fostering a holistic understanding of their synergies.

Dr. May is an active member of the national and international scientific community in Engineering Education Research. He has also organized several international conferences himself – such as the annual "International Conference on Smart Technologies & Education (STE)" – and serves as a board member for further conferences in this domain and for several Divisions within the American Society for Engineering Education.

KICK 4.0 - AI Chatting Skills in the Engineering Laboratory

Introduction

The KICK 4.0 project explores the use of language model-based (LLM) generative AI (genAI) technologies in laboratory-based engineering education in order to improve learning outcomes of students in the laboratory course on the one hand and to raise students' awareness of responsible use of genAI on the other. The aim is to make it possible to experience the opportunities and limitations of these AI technologies in university teaching, both from the perspective of instructors and students. Both target groups should significantly improve their skills in dealing with the relevant technologies based on their own experiences. To this end, pedagogical scenarios for real-world problem solving are developed for fluid mechanics labs, where students use genAI for research assignments. This approach aims to provide students with essential genAI skills for society and the world of work in a reflective and critical way and to promote their problem-solving and creative thinking skills. The stated objectives describe the specific problems that the project addresses and these can be summarized in four research questions:

RQ1) What skills are required when dealing with LLM-based generative AI tools?
RQ2) What are the strengths and limitations of LLM-based generative AI tools?
RQ3) How can users be supported in the development of individual problem-solving skills and creative thinking?
RQ4) What do scenarios for the effective use of LLM-based generative AI tools technologies in university laboratory teaching look like?

This paper addresses RQ4 by describing the development of criteria for evaluating the effectiveness of a RAG chatbot and its implementation in a fluid mechanics course offered by the Chair of Fluid Mechanics (University of Wuppertal), where students learn to use complex software to solve fluid mechanics problems.

Method

The KICK 4.0 project is part of the fluid mechanics module in the mechanical engineering program at the University of Wuppertal. This includes the Computational Fluid Dynamics (CFD) laboratory course. In this course, students learn how to use complex software (OpenFOAM) to solve fluid mechanics problems. The implementation of the project in an ongoing course and the goal of involving students and instructors in the development process suggest an implementation based on a Design-Based Research (DBR) approach.
In the context of a DBR approach, the term intervention describes targeted measures or changes within an educational context that aim to optimize learning processes, test new didactic concepts, or overcome existing challenges. Interventions should only be evaluated summatively after further refinement, i.e., the development potential of the interventions should be exhausted before an approach is rejected as unsuitable [1].

The DBR approach attempts to combine the development of innovative solutions to practical educational problems with the acquisition of scientific knowledge [2] and follows a cyclical and iterative process in which design, testing, analysis, and redesign continuously build on each other. On the one hand, this increases the quality of innovations in teaching and learning research. On the other hand, provides relevant insights for the specific field of practice [3]. The core idea of DBR is to examine learning situations not in isolated laboratory environments, but in real-life situations [4]. The objectives pursued are always twofold: On the one hand, relevant problems from educational practice should be solved, and on the other hand, theory-developing or theory-generating results should be achieved [5].

In the KICK 4.0 project, the specific problems have already been formulated (see research questions). The DBR approach can therefore be divided into four phases, which are briefly described below (see Figure 1).

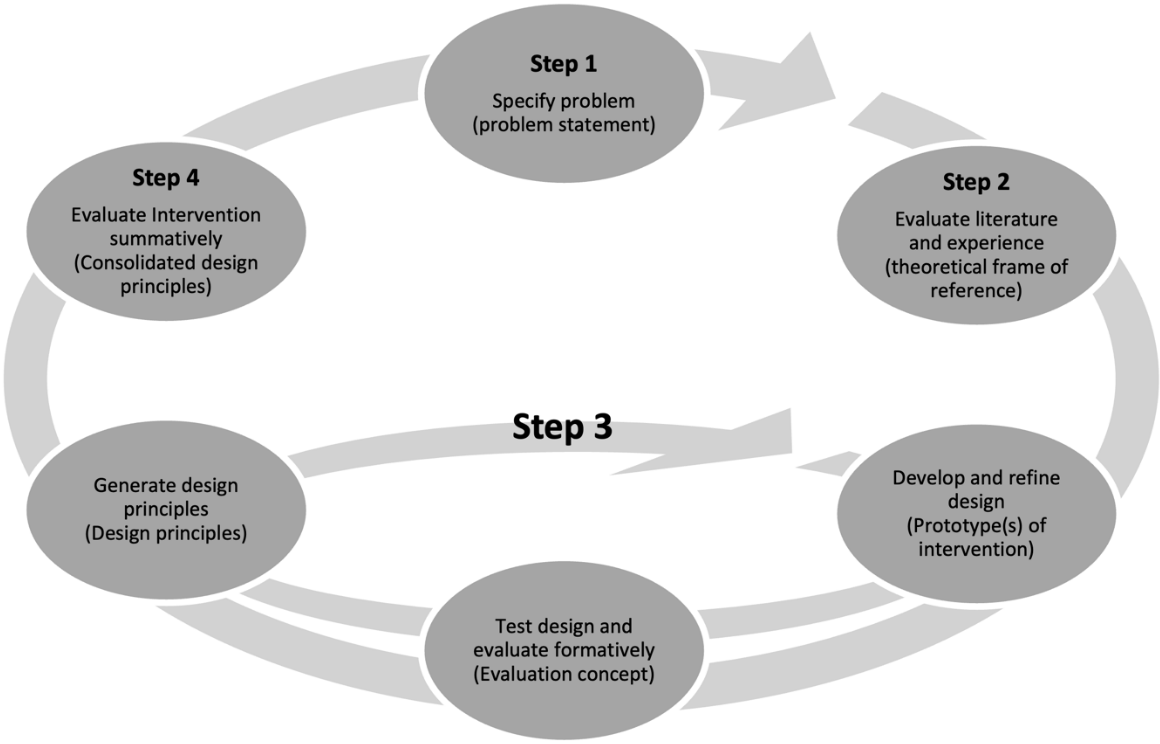

Figure 1: Adapted research and development cycle according to [1]

The first step is to conduct a requirements analysis. This involves analyzing a practical problem in a theory-driven manner, e.g., by determining the current state of research in a systematic literature review or by determining the status quo by means of an initial survey [6]. The use of surveys illustrates that the DBR approach is user-oriented, as designs for concrete solutions to practical problems are developed in a theory-oriented process [4, 7, 8].

The second step consists of the theory-based development of various prototype designs, taking into account the results from step 1. At the same time, possible evaluation methods are developed to measure the effectiveness of the intervention measures developed.

In the third step, the intervention and evaluation phase, interventions are carried out in an iterative process using the developed designs, scientifically evaluated, and then revised. It is often advisable to first carry out the planned intervention with a single, small target group in an initial evaluation cycle [9]. The need for revision resulting from the analysis is then implemented and re-examined in adapted evaluation cycles. Through this cyclical process, the designs are continuously tested, refined, and evaluated with the involvement of all participant of people involved until, in the best case, a solution to the initially formulated problem is achieved through a design.

In the fourth step, design principles are derived and formulated from the developed solutions. At the same time, the underlying theory is further developed. This illustrates that DBR must not only be theoretically sound, but also contributes to the development of new theories or the further development of existing theories, while at the same time generating practical implications [10].

Design-Based Research Approach in KICK 4.0
The following steps describe the application of the DBR approach in the KICK 4.0 project. The definition and specification of the problem was carried out in the conception phase and summarized in the research questions in the introduction.

Step 1: Conduct Requirements analysis (evaluating literature and experience)
The requirements analysis comprises two aspects. The first part consists of a systematic literature review and document analysis on the current state of AI-based teaching in STEM subjects. A standardized nine-step procedure based on the approach of [11] and [12] was used for this purpose. The aim of the systematic literature review is to provide a comprehensive overview of LLM-based generative AI tools applications in laboratory-based engineering education.

The second part of the requirements analysis consists of surveying participant groups of people, taking into account the results of the literature review. The surveys will be conducted at the beginning and end of each semester. The target groups are students of the master's programs “Fluid Mechanics” and “Computer Simulation and Science” as well as university lecturers. Due to the expected low number of participants, qualitative methods of empirical educational research will be used to carry out, analyze, and evaluate the interventions. The Plus-Minus-Interesting (PMI) method developed by Edward de Bono offers the opportunity to address a problem, weigh up the pros and cons, and reflect on implications. The method helps to assess consequences and develop alternatives [13].

The second part of the survey consists of a focus group discussion on two specific questions:

- How can ChatGPT support you as a student in the CFD course?
- In your opinion, are there use cases in the context of the CFD course where the use of ChatGPT is particularly suitable?

The surveys are evaluated using qualitative content analysis based on Mayring’s approach [14] with inductive category formation. Similar to prototype development, the evaluation methods used to measure the effectiveness of the prototypes are also subject to an iterative process and are further developed in each development loop.

In the first iteration loop, the students' task is to use ChatGPT in the course as LLM-based generative AI tool without influence from the instructors. The aim is to determine what expectations the groups have of LLM-based generative AI tools and what challenges and potential they see in its use in laboratory teaching. In the following iteration loops, the evaluation methods are adapted to the development of the prototypes.

Step 2: Development and refinement of the prototype
Based on the results of the requirements analysis, a prototype will be developed for use in a real-life scenario for the effective application of LLM-based generative AI. In addition, evaluation criteria will be derived from the results in order to assess the effectiveness and practicability of the prototype. Further interviews with the relevant groups of people will help to improve the defined criteria, validate the effectiveness of the prototype, and develop optimization options where necessary.

Step 3: Evaluate and redesign the intervention
This loop will be repeated in the following semester, the prototype will be revised, and the result will be further refined.

Step 4: Derive design principles
Based on the results, design principles will be developed that highlight possible applications of LLM-based generative AI technologies in laboratory-based STEM engineering education. This is to ensure that the results are transferable to other engineering disciplines and contribute to the dissemination of the concept. The project results will be published and made available to the community for further research.

Procedure for qualitative content analysis
Qualitative content analysis involves summarizing the text and presenting the meaning contained in the text in categories, which are in turn organized into a system. The category system, with categories, subcategories, category definitions, and anchor examples, represents the latent meaning contained in the evaluated texts [15].
The process model for qualitative content analysis, based on Mayring's approach, comprises eight steps (see Figure 2). The aim of content analysis is to "obtain the essential content [...] and, through abstraction, create a manageable corpus that still reflects the source material" [14].

Step 1: Determination of the material.
The question here is: Which material should be analyzed? Or, for example, should a selection of the material be made in order to answer only a specific question?
Step 2: Analysis of the situation in which the material was created.
How was the material created? Who wrote the text? Does the cognitive or sociocultural background play a role?
Step 3: Analyze the formal characteristics of the material.
Here, it is important to examine the form in which the material is available. A written text or transcription is usually required for content analysis.
Step 4: Determining the direction of the analysis.
Here, it is important to define the focus of interpretation. Content analysis can, for example, target emotional and cognitive content or the author's intentions.
Step 5: Theory-based differentiation of the research question.
In this step, a research question should be defined according to which the material is to be examined.
Step 6: Determining the analysis technique.
Mayring suggests three analysis techniques: summarization, explication, or structuring (see Step 7). In this step, you select which one to use, or whether to use a combination [14].
Step 7: Performing the analysis.
One or more of the techniques of summarization, explication, or structuring are now applied to analyze the material. The focus is always on developing a category system. These categories are derived from the theory either inductively or deductively [14].
Summary (inductive category formation): The aim of the analysis is to reduce the material while retaining the essential content. Selection criteria are defined in advance based on the research question and literature analysis. The text material is then read to familiarize oneself with the content. With the selection criteria in mind, the material is processed line by line. The content-bearing passages are paraphrased and these paraphrases are summarized into categories in a multi-stage process. The category designation is a term or sentence that often comes from the text. If text passages with similar meanings are found in the following, they are assigned to the category (subsumption). If new text passages are found that cannot be assigned to any of the categories formed, a new category is created. Once 10 to 50% of the material has been processed and no new categories are being formed, the category system is revised.

This involves checking whether the chosen level of abstraction corresponds to the text (i.e., the content is summarized, but the summary does not result in too much loss of meaning). If the category system is changed, the previously coded material must be reviewed again. Ultimately, all text passages should be assigned to categories. This is followed by the interpretation of the category system: answering the questions with the help of the categories and on the basis of the theory. This technique most closely corresponds to an inductive approach [16].
Explication: Explication focuses on clarifying incomprehensible or discrepant passages in the text. Additional material from neighboring text passages or additional sources is used to understand the text passages [14].
Structuring (deductive category formation): Certain aspects or a specific structure should be filtered out of the material, e.g., a typification should be carried out. The aim is to create a category system in which each text passage is classified, thereby capturing the structure of the material. This is done by defining categories, anchor examples, and coding rules. Anchor examples are specific text passages that describe the category in a prototypical way. Coding rules help to clearly assign text passages. The category system is applied and revised based on the material. Here, too, an evaluation and interpretation follow using the finished system based on the questions and theory [14]. With the help of the classification of text passages into the category system, individual representations are thus generalized across cases [14].
Step 8: Interpretation of the results.

Figure 2: Procedure according to Mayring [14]

Structure of the RAG system

A *Retrieval-Augmented Generation (RAG) system consists* of several components that work together to answer a query by incorporating external knowledge (see Figure 3). The process is as follows: A user query is first prepared for the search. Then, a retrieval module is used to search for relevant information in an external knowledge base. This retrieved information is then passed on, together with the original query, to a generator module (an LLM), which formulates the final, enriched answer [18].

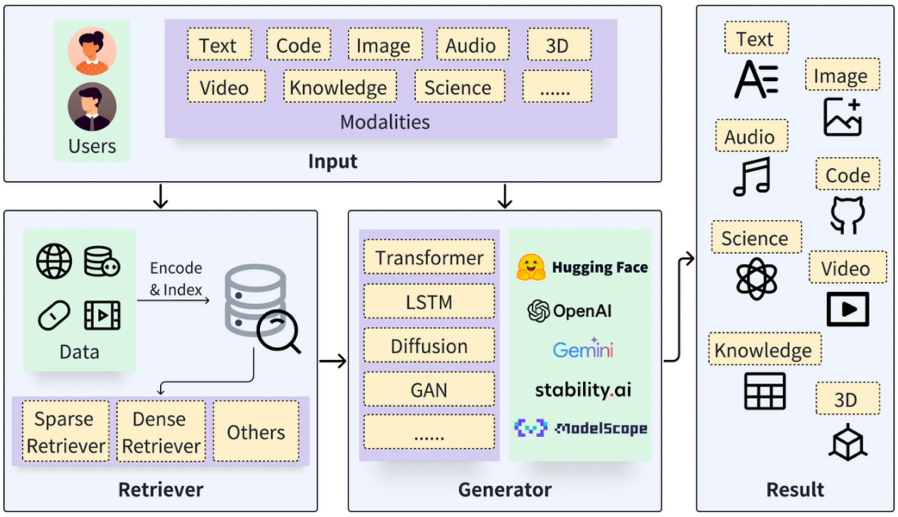

Figure 3: Schematic structure of a RAG system [17]

The RAG system retrieves the information from the *knowledge base*, an external data source. This can be, for example, texts of various kinds, images, or structured data from databases (SQL, NoSQL) [19]. The quality and timeliness of the knowledge base are crucial for the performance of the RAG system. This data must be pre-processed for use in a RAG system. This includes removing artifacts and chunking, i.e., dividing the content into short text passages (chunks). These chunks are converted into a searchable form and stored in a vector index [20].

The *retrieval module* identifies the most relevant chunks from the knowledge base based on the user query. It thus forms the bridge between a query and external content. Since a large amount of external content leads to a large number of chunks, approximate nearest neighbor (ANN) algorithms such as HNSW are often used [21].

The *generator module* is usually a large language model and has the function of formulating the response for the user. It contains the original user query, the most relevant chunks, and, if necessary, additional instructions via system prompts [22].

Results Requirements analysis

Qualitative content analysis according to Mayring

Step 1: The material used consists of the written results of the workshops conducted by the participating groups. As part of a preliminary study, students were invited to use ChatGPT as a digital assistant during the CFD lab course. The chat logs allow conclusions to be drawn about the correctness of the chatbot's answers, but do not provide any information that answers the research questions in the project (see Introduction) and are therefore not included in the analysis.

Step 2: The material was created during the workshops. All workshop results were recorded in writing on paper and presented by the participants. The groups consisted of either students or instructors.

Step 3: All results are available in written form and were digitized in advance.

Step 4: The survey was conducted with the participation of instructors and students. Therefore, all statements were divided into the categories “written by students” and “written by instructors” for the purposes of evaluation. The statements were also examined to determine the extent to which they affected students, instructors, or both groups. One example is the point “time savings.” This point is relevant for both groups. For students, it is particularly relevant for organizing their own studies. For instructors, it is particularly relevant for preparing (and also following up on) their own courses. Both groups benefit in the same way, but in different application scenarios.

Step 5: In order to define the research question, it is necessary to look at the project as a whole. The overarching goal is to investigate possible applications of AI tools in laboratory-based engineering education, specifically incorporating the perspectives of students and instructors. In addition, the priorities of the groups of people should provide information on criteria that can be used to evaluate the effectiveness of the chatbot. Therefore, the research question for the qualitative content analysis is:

“What priorities do users set when using AI tools, and what conclusions can be drawn from this about the effectiveness of the object of investigation (chatbot)?”

Step 6: In inductive category formation, categories are derived from the material under investigation and gradually generalized. In deductive category formation, conclusions are drawn from a general theory or rule to specific individual cases.

Since the participants' data should be analyzed impartially, the inductive approach was chosen.

Step 7: Implementation: The texts were grouped and summarized according to their influence on students and instructors. In the second round, similar formulations, e.g., the use of synonyms, were combined into one formulation. In the third round, initial categories were formulated from the responses. This process was carried out iteratively until no further summarization was possible. In some cases, subcategories were also created for individual categories. For example, the subcategories “finances,” “time,” and ‘resources’ were created for the category “availability of genAI tools”.

Step 8: The results of the analysis are the categories: competencies, availability, personalized learning, summaries, perspectives, social behavior, credibility, work organization, and ethics. These criteria were used in conjunction with the results of the literature review to develop the “Criteria for the perceived usefulness of genAI applications in a course context” (Short “perceived usefulness criteria” (see Table 1).

Table 1: Results of the requirements analysis

	Criteria for the perceived usefulness of genAI applications in a course context	Target groups
1	Allow user to experience system boundaries	students and instructors
2	Provide credibility of results	students and instructors
3	Availability of LLM-based genAI tools	students and instructors
4	Address ethical considerations	students and instructors
5	Feedback that promotes learning management and content comprehension	students
6	Receive adaptive feedback	students
7	Support instructional processes	instructors
8	Provide Transparency and traceability of results	instructors

Requirements analysis.

The results of the requirements analysis (Step 1 of the DBR Approach in KICK 4.0) consist of (a) literature review findings, and (b) qualitative content analysis results. Due to space constraints, the detailed results of the systematic literature review and the qualitative content analysis will be presented in separate publications. These results led to the "perceived usefulness criteria" and are summarized in Table 1. These criteria serve, on the one hand, as a framework for the development of the chatbot and, on the other hand, for the development of the evaluation design.

Point 1: "Allow user to experience system boundaries"

Addresses RQ 2. Users should learn about the potential as well as the limitations of generative AI tools. The learning activities are designed to expose boundary cases so that students can recognize limitations and develop verification habits.

Point 2: "Provide credibility of results"

Refers to statements made by students and instructors. Both groups have reservations about the content of the responses generated by AI tools. Users should be able to form their own judgment regarding the tool's credibility by learning prompting strategies and experiencing "boundary cases".

Point 3: "Availability of genAI tools"

Refers to the fact that AI tools require financial and material resources. Paid versions are generally significantly more powerful, thereby creating an imbalance between users and non-users. Access to hardware or fast and stable internet connections can also play a role.

Point 4: "Address ethical considerations"

Refer to statements from students who feel they are at a disadvantage because they do not want to or cannot use AI tools, and therefore believe they have fewer chances of successfully completing their studies. If the use of AI during one's studies leads to better grades than studying without AI support, this could spark a debate about fairness in higher education. The question of how to design exam formats, such as term papers, would then also need to be re-examined.

Point 5: "Feedback that promotes learning management and content comprehension"

Students and instructors see great potential in support for the learning process and in the individual structuring of their studies. Students also have expectations regarding feedback from AI tools. The AI tool is designed to provide answers and guidance that support the learning process, not ready-made answers.

Point 6: "Receive adaptive feedback"

The effectiveness of feedback depends on the recipient's prior knowledge. For those with less prior knowledge, there is a risk that overly detailed and extensive feedback may make them feel uncertain. Conversely, those with extensive background knowledge may become bored by feedback that is too superficial.

Point 7: "Support the instructional process."

This point is addressed to educators. GenAI tools are seen as having the potential to make the organization of the teaching process more effective. Possible future applications could include the analysis of exam results or the creation of teaching materials.

Point 8: "Provide transparency and traceability of results"

It is particularly important for instructors to know which sources the AI tool has access to and how these results are generated.

Based on the "perceived usefulness criteria", three approaches were derived to implement these criteria within the KICK 4.0 project.

Approach 1: Addresses criteria 3, 4, 7, and 8. This involves implementing a local chatbot that has access to predefined documents and is hosted on a local server. The original plan was to use LLaMA as the LLM tool. During the development of our own local RAG system, it became clear that implementing such a chatbot requires not only time but also a high level of programming expertise. The tool must also be able to compete with the constantly evolving offerings on the market (e.g., ChatGPT, Claude, Gemini). Since this is not feasible with the

available human and technical resources, the focus has shifted to approaches 2 and 3, and the implementation of LLaMA has been abandoned.
Approach 2: Addresses criteria 1 and 2 and describes the creation of a prompting guide. This prompting guide first outlines basic prompting strategies to help students understand how to create prompts that yield high-quality outputs. The second objective of the Prompting Guide is to encourage students to interact with the AI tool on a topic they are very familiar with. This is intended to help them draw on their own expertise to assess the quality of the chatbot's responses. Ultimately, the Prompting Guide is intended to lead to better results, a more accurate assessment of the quality of the responses, and increased confidence in the capabilities and limitations of the AI tool.
Approach 3: Addresses criteria 5 and 6 and describes how the AI tool's feedback is controlled via system prompts to generate feedback that promotes learning and is tailored to the individual. The AI tool's feedback is intended to fulfill two key tasks. First, depending on the context (small talk, short definition, technical question, OpenFOAM), the KICK chatbot's responses should support problem-solving without providing the solution in its entirety. The second task involves analyzing the user's prompt in terms of technical terminology, the specificity of the question, and spelling. The goal is to use this analysis to infer the user's prior knowledge.

Implementation of the RAG System – Chatbot KICK 4.0 – and preliminary results
Didactic scenario structure. The course is conducted in an in-person format, with each student working on their own PC with internet access. To support the individual learning process, the chatbot is integrated into the course as a "personal digital assistant" and is available to students throughout the entire course. Access is provided on an individual basis as needed, allowing students to use the chatbot as needed to clarify questions of understanding, reflect on their own approaches to solutions, and receive support in completing assignments.

Implementation of the RAG system. The RAG chatbot we developed combines a lexical retriever based on BM25, the Weaviate vector database as the retrieval infrastructure, and ChatGPT as the generative model for final response generation. Lecture notes on fluid mechanics and CFD and a manual on the use of OpenFOAM serve as sources of information. The segmentation of the documents into semantically coherent text sections, the chunks, was set to 500 words with an overlap of about 50 words between consecutive chunks. In the retrieval step, the query is passed to Weaviate, where a lexical retrieval using BM25 is performed. BM25 provides a ranking of those chunks with the highest lexical relevance to the query. The selection is made using a keyword list. The highest-ranked chunks are merged into a structured prompt. In addition, system prompts influence the response behavior. In order to provide feedback that promotes learning the system prompt is designed to elicit learning-promoting feedback.

Preliminary results: Comparison of the KICK Chatbot vs. GPT-5 (ChatGPT)
To evaluate the performance of the KICK chatbot in a preliminary comparison with ChatGPT, a question set containing 15 course-related questions about the OpenFOAM software was created as part of the CFD course. The questionnaire is based on posts from a course-related Q&A forum as well as on ChatGPT logs from the first round of the requirements analysis (see Step 1, DBR approach). The chatbot performed well on debugging tasks in the OpenFOAM context and on questions regarding the theoretical background. Its performance was weaker when it came to deriving concrete, practical solutions from theoretical questions.
Figure 4 shows the results of the comparison. GPT-5 (ChatGPT) provided a response in all test cases; 93% of these responses were rated as correct. The KICK chatbot, on the other hand, aims to refrain from providing a response when there is insufficient evidence. In the

test, 40% of the responses were rated as correct and 60% as non-relevant. Among the non-relevant cases, 7% consisted of responses that addressed a different question.

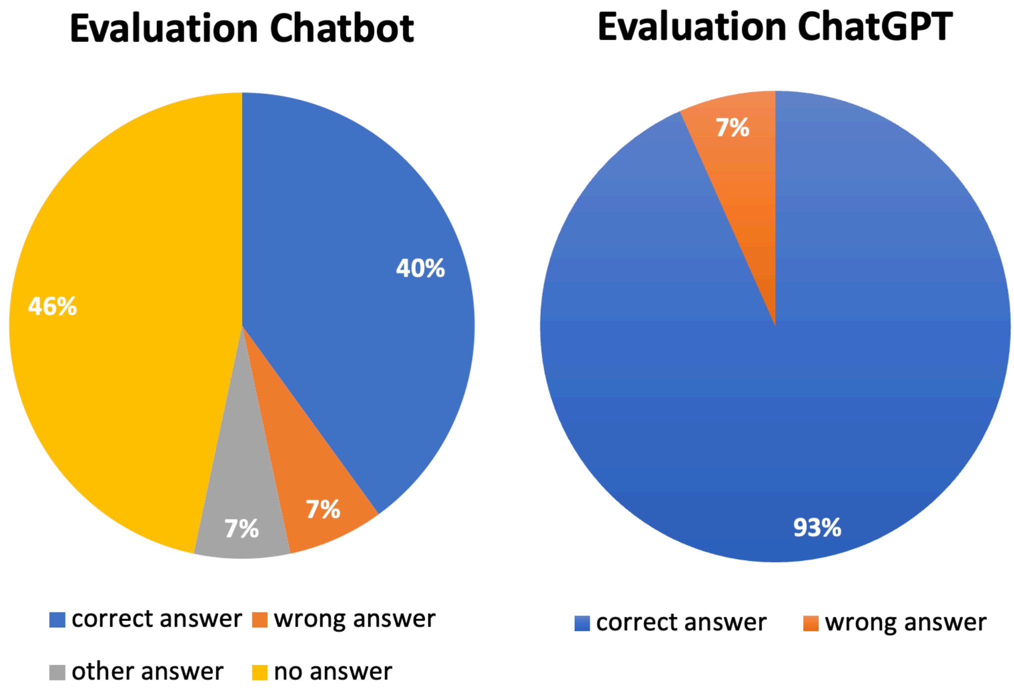

Figure 4: Comparison of the Results between KICK Chatbot and GPT-5 (ChatGPT)

Discussion and further steps

Limitations.

To minimize computational requirements and thus ensure a quick response from the KICK chatbot, user prompts were divided into four categories (small talk, short definition, technical question, OpenFOAM) based on keyword lists. Different sources were assigned to these categories. Managing the keyword lists has proven cumbersome over several iterations, as, for example, assigning a prompt to the wrong category due to missing or incomplete keywords could lead to poor results. The programming effort required to develop and improve a RAG system is considerable. For this reason, the approach of using a local system and implementing LLaMA will not be pursued further. Another consequence is that it is not possible to test how instructors would use a locally hosted chatbot/LLM.

Next Steps.

Four areas will be further developed within the project.

(1) A prompting guide to introduce students to prompting strategies. Another component of the guide is designed to enable students to explore the limitations of AI tools.

(2) Further development of the KICK chatbot at the level of system prompts. The aim is, on the one hand, to generate supportive feedback and, on the other hand, to generate adaptive feedback based on indicators in the user's prompt.

(3) Evaluation of the implemented measures. The revised KICK chatbot will be made available to students during the course. In a pre-test, students provide information about their prior experiences and habits with AI chatbots, as well as their prompting behavior. They are then given an introduction to using AI chatbots and a prompting guide.

After completing the course, students will be asked about changes in their prompting behavior and their perceptions based on the feedback they received from the KICK chatbot.

(4) The results will be incorporated into the formulation of the design principles.

References

[1] Euler, D., (2014). *Design-Research – a paradigm under development*. In D. Euler & P.F.E. Sloane (Hrsg.). Design-Based Research (15-44). Stuttgart: Franz Steiner Verlag.

[2] Shavelson, R. J., Phillips, D.C., Towne, L. & Feuer, M. J. (2003). *On the science of education design studies*. Educational Researcher, 32 (1), 25-28

[3] Klees, G. & Tillmann, A. (2015): *Design-Based Research als Forschungsansatz in der Fachdidaktik Biologie*. Journal für Didaktik der Biowissenschaften (F) 6, (2015), S. 92-110

[4] Reinmann, G. (2005). *Innovation ohne Forschung? Ein Plädoyer für Design-Based Research-Ansatz in der Lehr-Lernforschung*. Unterrichtswissenschaft 33(1), 52-69.

[5] Lehmann-Wermser, A. & Konrad, U. (2016). *Design-Based Research als eine der Praxis verpflichtete, theoretisch fundierte Methode der Unterrichtsforschung und -entwicklung. Methodologische Grundlagen, dargestellt am Beispiel eines Forschungsprojekts im Bandklassen-Unterricht*. In: Knigge, J., Niessen, A. (Hrsg.): Musikpädagogik und Erziehungswissenschaft. Münster; New York: Waxmann 2016, S. 265-280.

[6] Schmiedebach, M., Wegner, C. (2021). Design-Based Research als Ansatz zur Lösung praxisrelevanter Probleme in der fachdidaktischen Forschung. In *Bildungsforschung*, (2021) 2, S. 1-10. DOI: https://doi.org/10.25656/01:23920

[7] Knogler, M. & Lewalter, D. (2014). Design-Based Research im naturwissenschaftlichen Unterricht – Das motivationsfördernde Potential situierter Lernumgebungen im Fokus. *Psychologie in Erziehung und Unterricht*, 61, 2-14.

[8] Plomp, T. (2013). Educational Design Research: An Introduction. In: T. Plomp & N. Nieveen (Hrsg.), Educational Design Research, Part A. Enschede, SLO, 10-51.

[9] (Scheersoi & Tessartz 2019, 2) Scheersoi, A. & Tessartz, A. (2019): Design-Based Research – ganz praktisch. In *Bildungsforschung*. DOI: https://doi.org/10.25539/bildungsforschun.v0i1.283

[10] Ruthven, K., Laborde, C., Leach, J. & Tiberghien, A. (2009). *Design Tools in didactical research: Instrumenting the epistemological and cognitive aspects of the design of teaching sequences*. Educational Researcher, 38(5), 329-342

[11] Borrego, M., Foster, M. J., Jeffrey E. Froyd, J. E. (2014) *Systematic Literature Reviews in Engineering Education and Other Developing Interdisciplinary Fields*. In: Journal of Engineering Education VC 2014 ASEE. http://wileyonlinelibrary.com/journal/jee, Vol. 103, No. 1, pp. 45–76

[12] Heil, E., A. (2021). *Methode der systematischen Literaturrecherche. Justus-Liebig-Universität Gießen*. DOI: http://dx.doi.org/10.22029/jlupub-17189

[13] Becker, J., H., Ebert, H., Pastoors, S.: *Praxishandbuch Berufliche Schlüsselkompetenzen*, Springer (2018).

[14] Mayring, P. (2007). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*, Weinheim.

[15] Mayring, P. (2002). *Einführung in die qualitative Sozialforschung: eine Anleitung zum qualitativen Denken*, Weinheim.

[16] Kohlbacher, F. (2006). *The use of qualitative content analysis in case study research*. In: Forum Qualitative Sozialforschung 7. Artikel 21.

[17] Zhao, P., Zhang, H., Yu, Q. et al. *Retrieval-Augmented Generation for AI-Generated Content: A Survey*. Data Sci. Eng. (2026). https://doi.org/10.1007/s41019-025-00335-5

[18] Lewis, P., et al.: *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. (2021). http://arxiv.org/abs/2005.11401. arXiv: 2005.11401.

[19] Ai, Q., et al.: „*Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community*“. In: AI Open 4. (2023).

https://linkinghub.elsevier.com/retrieve/pii/S2666651023000049, S. 80–90. issn: 26666510. DOI: 10.1016/j.aiopen.2023.08.001.

[20] Roucher, A.: A*dvanced RAG on Hugging Face Documentation Using LangChain*. (2025). https://huggingface.co/learn/cookbook/advanced_rag.

[21] Glass, M.:l et al.: *Re2G: Retrieve, Rerank, Generate*. (2022). http://arxiv.org/abs/2207.06300. DOI: 10.48550/arXiv.2207.06300. arXiv: 2207.06300 [cs].

[22] Lewis, P., et al.: *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. (2021). http://arxiv.org/abs/2005.11401. arXiv: 2005.11401.