A Qualitative Evaluation of an AI-Supported Quiz Application to Assess Learning Progress

Betiel Woldai, University of Applied Science Ansbach, Germany
Sophie Henne, University of Applied Science Ansbach, Germany
Mascha-Lea Fersch, University of Applied Science Ansbach, Germany
Sudarshan Kamath Barkur, University of Applied Science Ansbach, Germany
Sigurd Schacht, University of Applied Science Ansbach, Germany

The Paris Conference on Education 2023
Official Conference Proceedings

Abstract

In a current research project at the Ansbach University of Applied Science, an AI-based quiz function was created to serve as a voluntary student-oriented support offer to determine their learning progress in their respective courses by means of conducting self-assessment quizzes. The application takes lecture scripts as input and applies a question generation model to create questions that students can answer. In order to evaluate the given answers, another language model is involved to perform Natural Language Inference (NLI). Users can engage with the system via a graphical user interface currently provided via a web app. To assess preliminary feasibility and perception of the model prototype, a qualitative focus group discussion following a semi-structured interview guideline prepared by the research team according to similar studies in the education field (Sek et al. 2012) was conducted with five participants. A transcript of the discussion was prepared and analyzed using the qualitative content analysis method according to Kuckartz. Overall, the quiz function was well received by the participants of the focus group. However, the prototype still has potential when it comes to generating meaningful questions and transparently assigning categories to the given answers. Furthermore, the quiz parameters should be individually adjustable by users. In the following paper, the development of the service is illustrated by outlining the considerations for the application design and the training procedure of the language models. Afterwards, the design of the qualitative focus group is described including the presentation of the results.

Keywords: Higher Education, Conversational AI, Learning Progress, Self-Assessment Quizzes, Digital Study Assistant

iafor

The International Academic Forum
www.iafor.org

1. Introduction

With the advent of artificial intelligence (AI) in the education sector, interesting potentials for its application at universities arise in particular in the field of Natural Language Processing. The goal of the current research project at Ansbach University of Applied Science is to develop a digital, intelligent assistant for study and teaching. The digital assistant will have four main application areas: a communicator component for answering questions, conversation, and mentoring; a planner to perform time management and course planning tasks; a motivator to actively manage learning success; and an analyzer, to provide the necessary information about the student's study and learning progress. The latest focus of the project is to establish an analyzing component that can be used to automatically measure learning progress of students.

Self-assessments carried out by the students are a suitable option for this purpose. Student self-assessment refers in general to a variety of mechanisms and techniques through which students assess and evaluate their own learning progress (Panadero et al., 2016). One form of self-assessments are online quizzes to test understanding of the course content (Bognár et al., 2021). Quizzes offer a dynamic environment due to their numerous customization options with regards to size, question types, grading, time limitations etc. (Gikandi et al., 2011). The score derived from answering questions provides immediate feedback to learners and thus supports them in monitoring whether they have achieved a learning objective or a desired level of performance in a course (Ćukušić et al., 2014).

Currently, however, the questions would still have to be developed manually by the respective lecturers. Large Language Models can support in this scenario as they are increasingly used for various language-based tasks like question-answering, text generation or summarization. Due to the possibility to enhance the capacities of transformer language models more and more, a trend towards increasing the scale of language models has emerged in recent years. Thus, the architectures of these models are no longer task-specific, but task-agnostic in design. Moreover, they are trained on large datasets that are also task-unspecific as well as domain-unspecific. (Wei et al., 2022). Trained once, LLM strongly perform in zero-, one- or few-shot settings at tasks defined on-the-fly like the automatic generation of questions (Brown et al., 2020). In addition, there are also verification mechanisms using Natural Language Inference to check the answers to questions for their correctness.

The following paper describes the development of such a system, which can be used for the automatic generation of questions for self-assessment quizzes as well as for the verification of the given answers. First, the application design of the quiz function including relevant findings from the literature are pointed out in section 2. Then, in section 3, the possible architecture of such a model is presented and the process for testing is described. Afterward, a qualitative focus group with students was conducted to assess preliminary feasibility and perception of the system prototype followed by a discussion of the results in section 6. Finally, a conclusion is drawn in section 7.

2. Application Design

This chapter describes the system that can be used for the automatic generation of quizzes to measure learning progress. For the development of the design, various studies were considered that use online quizzes for the self-assessment of students.

In a research project at the University of Stuttgart, a smartphone app was designed that offers students quizzes on modules of a selected course and provides direct feedback after answering the questions (display of an overview of the number of correctly and incorrectly answered questions). The quizzes can be used on a voluntary basis. In addition, the results are not stored and evaluated, but can only be viewed by the respective student (Pauli et al., 2020).

An application developed at the University of Graz also provides students with online quizzes for individual courses, however, via the learning management system (LMS) Moodle. Students are given the opportunity to voluntarily check their level of knowledge in a lecture based on questions about course material and automatically receive a grade for their results. The quizzes can be taken several times (Schweighofer et al., 2019).

In another example, which provides online quizzes for students via the LMS MyMathLab, the authors (Sek et al., 2012) point out various features that need to be considered when developing a quiz application. These include the number of questions, the number of attempts, the question format, time limitations, and the way the results are displayed.

Based on these examples, initial requirements for a system that can be used at Ansbach University of Applied Science were collected within the research team. Overall, the quiz function is intended to serve as a voluntary student-oriented support offer to determine their learning progress in their respective courses by means of conducting quizzes regardless of possible offers provided by lecturers. Accordingly, the application contains content-related questions about a course with a prompt evaluation of the given answers. The results are only visible to the respective student. In addition, the quizzes should be able to be conducted online and thus independent of location and time.

The use of language models for the development of the application enables an automatic generation of questions. Using lecture materials such as scripts or book extracts, questions about the content should be generated. In order to evaluate the given answers, another language model be involved to perform Natural Language Inference (NLI). The NLI model classifies the given answers into three classes with the respective probabilities and thus, statements about the correctness can be made. Based on these requirements, the development of a first prototype will be described in the following section.

3. Model Architecture

The proposed system consists of two components: (1) a model for question generation and (2) a model for NLI whereas the two models are not connected and work independently. First, questions are generated given a text passage. Then, answers provided by a user are checked for their accuracy. Users can engage with the system via a graphical user interface currently provided via a web app.

3.1 Question Generation Model

For the question generation task, a pre-trained T5 model is used as a foundation. Being trained on a huge amount of unlabeled data consisting of clean English text (the Colossal Clean Crawled Corpus), the T5 model is able to perform various Natural Language Processing (NLP) downstream tasks e.g. summarization or translation (Raffel et al., 2020). Its architecture is based on an encoder-decoder transformer implementation which converts all NLP problems into a text-to-text format. In our case, a pre-trained T5-model fine-tuned on the GermanQuAD dataset which consists of extractive question and answer pairs was used (Dehio, 2022). Given a German text, the model generates a list of questions about it. As seen from others, beam search and top_k random sampling to generated a variety of questions was used (von Platen, 2020).

3.2 Natural Language Inference

To reach the goal of verifying if given answers to a question are correct NLI is used. Natural language inference is the task of determining whether a “hypothesis” is true (entailment), false (contradiction) or undetermined (neutral) given a “premise” (Lokshyn, 2022). A pre-trained mDeBERTa-v3-base model was used trained on a large multilingual dataset containing NLI hypothesis-premise pairs (Laurer, 2022). The transformer-based architecture is able to compare a given answer with the underlying text and return a score with regards to the tree classes entailment, contradiction and neutral.

3.3 Graphical User Interface

Users can access the system via a web app that is created with the Streamlit library. Multiple user interfaces have been build. The landing page displays a menu for selecting the quiz topic (see Fig. 1).

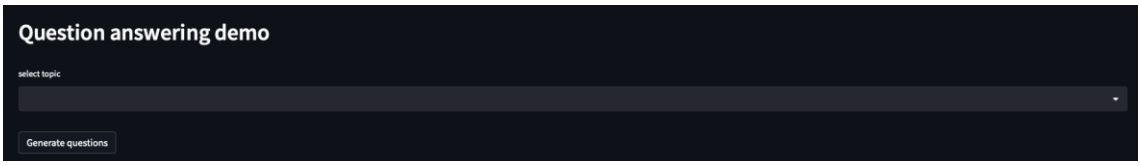
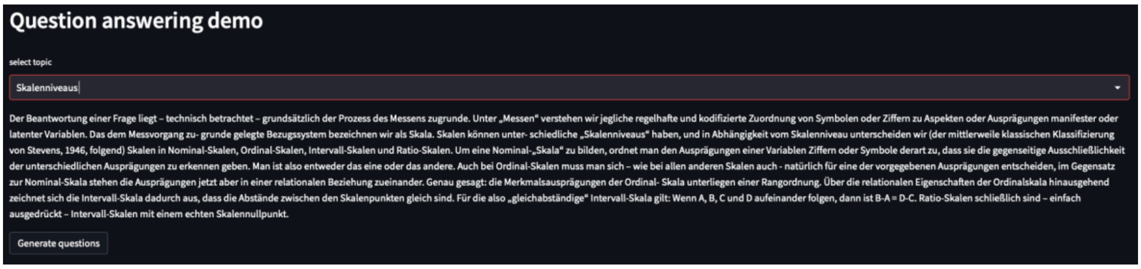


Figure 1: Landing page of the quiz function.

After selecting the topic, the user is automatically redirected to the second interface displaying the underlying text on the basis of which the associated questions are generated (see Fig. 2).



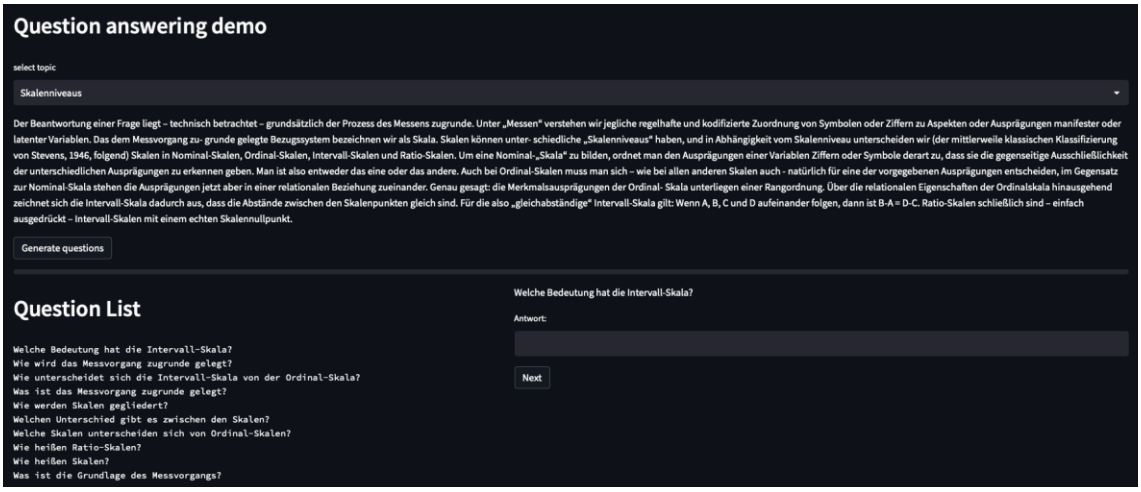
select topic

Skalenniveaus

Der Beantwortung einer Frage liegt – technisch betrachtet – grundsätzlich der Prozess des Messens zugrunde. Unter „Messen“ verstehen wir jegliche regelhafte und kodifizierte Zuordnung von Symbolen oder Ziffern zu Aspekten oder Ausprägungen manifester oder latenter Variablen. Das dem Messvorgang zu- grunde gelegte Bezugssystem bezeichnen wir als Skala. Skalen können unter- schiedliche „Skalenniveaus“ haben, und in Abhängigkeit vom Skalenniveau unterscheiden wir (der mittlerweile klassischen Klassifizierung von Stevens, 1946, folgend) Skalen in Nominal-Skalen, Ordinal-Skalen, Intervall-Skalen und Ratio-Skalen. Um eine Nominal-„Skala“ zu bilden, ordnet man den Ausprägungen einer Variablen Ziffern oder Symbole derart zu, dass sie die gegenseitige Ausschließlichkeit der unterschiedlichen Ausprägungen zu erkennen geben. Man ist also entweder das eine oder das andere. Auch bei Ordinal-Skalen muss man sich – wie bei allen anderen Skalen auch - natürlich für eine der vorgegebenen Ausprägungen entscheiden, im Gegensatz zur Nominal-Skala stehen die Ausprägungen jetzt aber in einer relationalen Beziehung zueinander. Genau gesagt: die Merkmalsausprägungen der Ordinal- Skala unterliegen einer Rangordnung. Über die relationalen Eigenschaften der Ordinalskala hinausgehend zeichnet sich die Intervall-Skala dadurch aus, dass die Abstände zwischen den Skalenpunkten gleich sind. Für die also „gleichabständige“ Intervall-Skala gilt: Wenn A, B, C und D aufeinander folgen, dann ist B-A = D-C. Ratio-Skalen schließlich sind – einfach ausgedrückt – Intervall-Skalen mit einem echten Skalennullpunkt.

Figure 2: Display of the text used to generate the questions.

The "Generate questions" button outputs a list of 10 questions. In addition, the questions appear individually above the corresponding input field for typing in the answers (see Fig. 3).



Question answering demo

select topic

Skalenniveaus

Der Beantwortung einer Frage liegt – technisch betrachtet – grundsätzlich der Prozess des Messens zugrunde. Unter „Messen“ verstehen wir jegliche regelhafte und kodifizierte Zuordnung von Symbolen oder Ziffern zu Aspekten oder Ausprägungen manifester oder latenter Variablen. Das dem Messvorgang zu- grunde gelegte Bezugssystem bezeichnen wir als Skala. Skalen können unter- schiedliche „Skalenniveaus“ haben, und in Abhängigkeit vom Skalenniveau unterscheiden wir (der mittlerweile klassischen Klassifizierung von Stevens, 1946, folgend) Skalen in Nominal-Skalen, Ordinal-Skalen, Intervall-Skalen und Ratio-Skalen. Um eine Nominal-„Skala“ zu bilden, ordnet man den Ausprägungen einer Variablen Ziffern oder Symbole derart zu, dass sie die gegenseitige Ausschließlichkeit der unterschiedlichen Ausprägungen zu erkennen geben. Man ist also entweder das eine oder das andere. Auch bei Ordinal-Skalen muss man sich – wie bei allen anderen Skalen auch - natürlich für eine der vorgegebenen Ausprägungen entscheiden, im Gegensatz zur Nominal-Skala stehen die Ausprägungen jetzt aber in einer relationalen Beziehung zueinander. Genau gesagt: die Merkmalsausprägungen der Ordinal- Skala unterliegen einer Rangordnung. Über die relationalen Eigenschaften der Ordinalskala hinausgehend zeichnet sich die Intervall-Skala dadurch aus, dass die Abstände zwischen den Skalenpunkten gleich sind. Für die also „gleichabständige“ Intervall-Skala gilt: Wenn A, B, C und D aufeinander folgen, dann ist B-A = D-C. Ratio-Skalen schließlich sind – einfach ausgedrückt – Intervall-Skalen mit einem echten Skalennullpunkt.

Question List

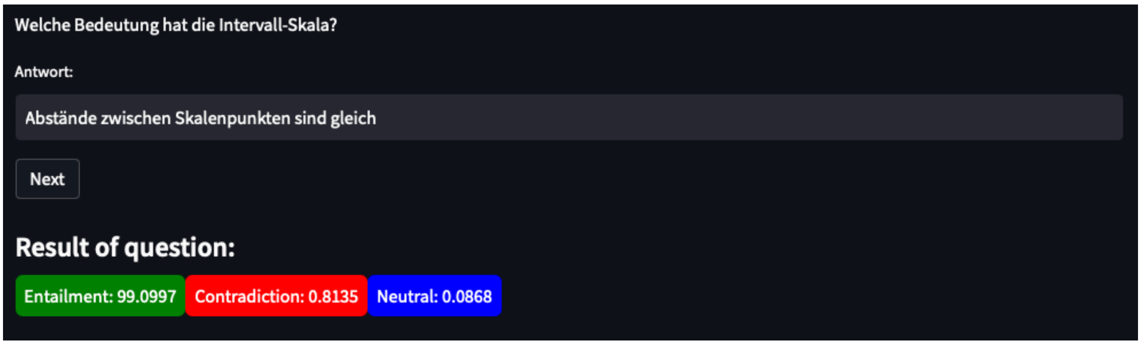
```
Welche Bedeutung hat die Intervall-Skala?
Wie wird das Messvorgang zugrunde gelegt?
Wie unterscheidet sich die Intervall-Skala von der Ordinal-Skala?
Was ist das Messvorgang zugrunde gelegt?
Wie werden Skalen gegliedert?
Welchen Unterschied gibt es zwischen den Skalen?
Welche Skalen unterscheiden sich von Ordinal-Skalen?
Wie heißen Ratio-Skalen?
Wie heißen Skalen?
Was ist die Grundlage des Messvorgangs?
```

Welche Bedeutung hat die Intervall-Skala?

Antwort:

Figure 3: Output of the question list.

After typing in the answer, via the "Next" button, the scores for the three classes are displayed below the answer field (see Fig. 4).



Welche Bedeutung hat die Intervall-Skala?

Antwort:

Abstände zwischen Skalenpunkten sind gleich

Result of question:

Entailment: 99.0997 Contradiction: 0.8135 Neutral: 0.0868

Figure 4: Answer verification.

4. Testing

As pre-trained models were used, no training but testing process was required. Both models were tested manually with a human in loop. The testing of the question generation model was based on several steps. In order to select a question-answering model a small test was conducted during which various pre-trained question generation models were used to check the quality of the generated question based on the same paragraph. Once a model was selected, it was detected that the paragraph length correlated with the repetition of questions. The bigger the paragraph, the better the quality of the questions generated without repetition. The generated questions were processed to remove the repeated questions by the model programmatically. For the NLI model the paragraphs with the generated questions were evaluated manually.

5. Evaluation

In order to evaluate the quiz function from the perspective of future users, it was decided to have an initial prototype tested by students during the development process. The goal of this evaluation was to get feedback on current features of the quiz function, but also to get ideas on how the function can be better adapted to the student’s needs. For this purpose, a qualitative focus group was conducted with several participants.

5.1 Participants

The participants of the study were approached through a university course and volunteered to participate in the focus group in December 2022. In total, 5 participants were recruited for the focus group consisting of three students, one professor and one scientific associate, all from the same study program. Among them there were four females and one male. Due to the limited research funds, there was no reimbursement for the focus group participation.

5.2 Data Collection

For the focus group, a semi-structured interview guideline was prepared by the research team. The semi-structured format was chosen to allow for open discussions. The questions were partly created based on prior literature research of similar studies in the educational field (Sek et al., 2012). The focus group was performed online via Zoom and moderated by two of the authors of this paper. As two specifications of the Analyzer were studied, the presentation and discussion was divided into two parts as well. In each part, the Analyzer component was first presented, after which the participants had the opportunity to test the function themselves followed by a discussion.

Question Category	Explanation	Example Question
Scope & Content	Questions concerning the scope and content quality of the questions	"Were the questions asked about the content understandable? "How did you feel about the number of questions in a section?"
Parameters	Questions concerning the different parameters of the model such as time restrictions	"Do you think there should be a time limit to answer the questions?" "When should the answers to the question be displayed?- Directly after the question or after completion of the quiz?" "Would it be helpful if you could give a due date for the quizzes on a particular topic?"
User Experience	Questions dealing with usability, navigation, and the output format	"Were you satisfied with the presentation of the results per question and overall?" "How did you perceive the navigation on the page? Did you know where to click to get the necessary information?" "How should the quiz ideally be delivered?"
Format	Questions dealing with the general output format	"What other possibilities do you know or use to check your level of learning progress?" "Would you prefer a different examination format? For example as a multiple choice test?"

Table 1: Semi-structured Interview Guideline.

5.3 Data Analysis

For data analysis we used the qualitative content analysis method (Kuckartz, 2016) in two stages according to (Schulz, 2012), since this procedure fits best to our research design. In a first step, the video recording was transcribed and reviewed several times by the research team. In a second step, the data were coded, categorized and after a final review adapted. The coding procedure was done using both an inductive and a deductive approach. As described previously the semi-structured interview guideline was based on three pre-defined categories. These categories were then adapted during the analysis, further categories were added and others renamed. The coding process was reviewed by other team members to maximize objectivity.

5.4 Results

In general, the quiz function was well received by the participants. Positive aspects that were mentioned cover the self-explanatory design of the system, the amount of questions and the display of a rating of the given answers. Participants were in general surprised and impressed with the capability of the question generation model.

In terms of the content quality of the generated questions, the feedback was mixed. While some of the questions encourage a deeper engagement with the topic, others were deemed less useful. For example, several times the same question was issued by the model only in a different wording. However, the structure of the questions was predominantly evaluated positively.

The natural language inference model did not fully meet the expectations of the participants. While the evaluation of the answers was initially emphasized favorably, the rating according to the three classes (entailment, contradiction or neutral) does not always appear to be plausible. Suggestions for improvement included, first, an explanation of how the response was classified, and second, the output of a sample solution.

Another intention behind the focus group was to obtain ideas regarding the format of the quizzes. Among other things, the addition of a time limit during the answering of the questions or the integration of a due date was pointed out. Limiting the number of attempts to complete a quiz could be another option. Regarding the answer format, multiple-choice quizzes were discussed as an alternative to the current open-question format.

Finally, the provision to the quiz function was discussed. In particular, students would welcome the offer of a smartphone app or integration via the learning management system used at the university in order to prevent additional media disruptions.

Category and sub-category	Definition	Quote (example)
1. Positive Aspects	All text passages that include positive aspects of the quiz function	“What I particularly liked about the quiz function is the structure, where you first read up on the topic and then have it in front of your eyes and directly answer various questions about it.” (Focusgroup2.2, paragraph 8)
2. Improvement Suggestions	All text passages that show improvement options for the feature	
2.1 Comprehensibility	All text passages that include suggestions for improvement on the comprehensibility of the result as well as information value of the questions	“A suggestion would also be to clarify what exactly "good" means, what "neutral" means, and why.” (Focusgroup2.2, paragraph 17)
2.2 User Experience / Navigation	All text passages that demonstrate how user experience and navigation were perceived	“Everything that is connected to authentication would be an additional hurdle” (Focusgroup2.2, paragraph 50)
2.3 Functionality	All text passages that demonstrate suggestions for improvement on various functions	
2.3.1 Time Limit	All text passages that demonstrate how a time limit can support the preparation for the exam	“I actually think it would be cool if you could choose whether you want to have a time limit or not. Because maybe if a topic is new, it would be great to have enough time to think about it. But if it's a topic that you've already had several times and are well-prepared for, you can say that you have a time limit, just like you would in an exam.” (Focusgroup2.2, paragraph 24)
2.3.2 Number of attempts	All text passages that demonstrate how limiting the number of attempts can support exam preparation	“Maybe it is also a point that can be left open, like limiting the time, because it simulates the feeling of an exam more strongly. But I generally don't think it's bad if it would be limited.” (Focusgroup2.2, paragraph 12)
2.3.3 Sample solutions	All text passages that demonstrate how a sample solution could improve the function	“Maybe it would be really cool if there was the possibility to display some kind of sample solution.” (Focusgroup2.2, paragraph 12)
2.3.4 Examination format	All text passages that identify various exam questions to support the learning process	“It depends, whether I am in the learning phase. I think, then I would prefer to actively engage with the material or is it just before the exam and I want to check where I stand. Then, hiding it is probably better.” (Focusgroup2.2, paragraph 37)
2.3.5 Due Date	All text passages that demonstrate how setting a due date for learning content can promote the learning process	“Everyone would find that helpful.” (Focusgroup2.2, paragraph 40)

Table 2: Results from the second part of the focus group, reflecting the testing of the quiz function. Texts have been translated from German.

6. Discussion

The objective of this focus group was to assess preliminary feasibility and perception of the quiz function. The findings suggest that especially the question generation model as well as the NLI model and the structure of the quizzes need to be adapted according to the needs of the future users. Since the question generation model does not yet offer a consistent quality of the created questions, further training of the model is necessary, possibly also on further data sets. The categorization of the NLI model must be made transparent to the user. This means that the underlying rules for the assignment of the classes would have to be displayed or hints would have to be given in order to answer a question completely correctly. Overall, more customization options are desired. Students would like more customization options, e.g., selecting answer formats, setting a time limit, setting the number of attempts to answer a question.

6.1 Limitations and Implications for Future Work

A shortcoming of the design of the focus group relates to the small sample size of the participant group. Moreover, the group was recruited from only one course and study program, which resulted in a low heterogeneity of the sample group. In contrast, this could also be seen as advantage, since one may assume a similar level of knowledge of the participants. Nevertheless, future focus groups should consider a larger, as well as more diverse participant base.

In terms of practical implications, the focus group resulted in several ideas for the further development of the quiz function, such as the proposal to integrate the application within the Learning Management System Moodle. The feature would be available in the same place as course materials. A benefit would be that Moodle offers a mobile application, with which students could access the quiz function on their smartphone. Another interesting aspect is the expansion of the model from students to teachers. For example, by supporting teachers in the creation of exam questions or the analysis of the learning progress of participants in a course.

7. Conclusion

In this paper, the use of language models to measure learning progress was assessed. The overall research goal was to develop an application that allows students to self-assess learning progress on course level. An AI-based quiz function was created using a question generation model and an NLI model. The application takes lecture scripts as an input and creates questions that can then be answered by students. The given answers are verified by the NLI model. A qualitative focus group was conducted to acquire insights about the application with regards to user experience, format, and content. Overall, the quiz function was well received by the participants of the focus group. However, the prototype still has potential when it comes to generating meaningful questions and transparently assigning categories to the given answers. Furthermore, the quiz parameters should be individually adjustable by users.

The study highlights the enormous potential that can be derived from the application of language models within the educational context. Moreover, practical implications for implementing an AI-based quiz function are described. The insights gained from the focus group will be incorporated in the further development of the quiz function.

References

Bognár, L., Fauszt, T., & Váraljai, M. (2021). The Impact of Online Quizzes on Student Success. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(11), 225. https://doi.org/10.3991/ijet.v16i11.21679

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Ćukušić, M., Garača, Ž., & Jadrić, M. (2014). Online self-assessment and students' success in higher education institutions. *Computers & Education*, *72*, 100–109. https://doi.org/10.1016/j.compedu.2013.10.018

Dehio, N. (2022). *German-qg-t5-e2e-quad*. https://huggingface.co/dehio/german-qg-t5-e2e-quad.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Kuckartz, U. (2016). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung* (3., überarbeitete Auflage). Beltz Juventa.

Laurer, M. (2022). *MDeBERTa-v3-base-xnli-multilingual-nli-2mil7*. https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

Lokshyn, O. (2022, Juni 9). Natural Language Inference: An Overview. *Towards Data Science*. https://towardsdatascience.com/natural-language-inference-an-overview-57c0eecf6517

Panadero, E., Brown, G. T. L., & Strijbos, J.-W. (2016). The Future of Student Self-Assessment: A Review of Known Unknowns and Potential Directions. *Educational Psychology Review*, *28*(4), 803–830. https://doi.org/10.1007/s10648-015-9350-2

Pauli, P., Koch, A., & Allgöwer, F. (2020). Smartphone Apps for Learning Progress and Course Revision. *IFAC-PapersOnLine*, *53*(2), 17368–17373. https://doi.org/10.1016/j.ifacol.2020.12.2088

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (arXiv:1910.10683). arXiv. http://arxiv.org/abs/1910.10683

Schulz, M. (2012). Quick and easy!? Fokusgruppen in der angewandten Sozialwissenschaft. In M. Schulz, B. Mack, & O. Renn (Hrsg.), *Fokusgruppen in der empirischen Sozialwissenschaft* (S. 9–22). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-19397-7_1

Schweighofer, J., Taraghi, B., & Ebner, M. (2019). Development of a Quiz – Implementation of a (Self-) Assessment Tool and its Integration in Moodle. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(23), 141. https://doi.org/10.3991/ijet.v14i23.11484

Sek, Y.-W., Law, C.-Y., Liew, T.-H., Bt Hisham, S., Lau, S.-H., & Pee, A. N. B. C. (2012). E-Assessment as a Self-Test Quiz Tool: The Setting Features and Formative Use. *Procedia - Social and Behavioral Sciences*, *65*, 737–742. https://doi.org/10.1016/j.sbspro.2012.11.192

von Platen, P. (2020, März). How to generate text: Using different decoding methods for language generation with Transformers. *Hugging Face*. https://huggingface.co/blog/how-to-generate

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. http://arxiv.org/abs/2206.07682

Contact email: b.woldai@hs-ansbach.de