

Leben in Deutschland: An Intelligent Tutoring Game for the German Citizenship Test

Felix Haase, Kemal Tanirak, Johannes Langer, Shamim Miroliaei, Jochen Mehlich and Ute Schmid

University of Bamberg, Germany

felix.haase@uni-bamberg.de

kemal.tanirak@stud.uni-bamberg.de

johannes.langer@uni-bamberg.de

shamim.miroliaei@uni-bamberg.de

jochen.mehlich@uni-bamberg.de

ute.schmid@uni-bamberg.de

Abstract: To apply for German citizenship, each candidate must complete a multiple-choice questionnaire of 33 questions with at least 17 correct answers. The official question database contains a total of 310 questions covering various topics such as history, politics, and values. While there are already courses, websites and apps to prepare for this test, there exists no digital serious game that creates an engaging environment and supports the learners individually. The aim of this proposed intelligent tutoring game “Leben in Deutschland” is to support learners in preparing for the test, by situating them in a German town, tracking question mastery in repeated trials, providing feedback on failed questions and allowing them to pose open questions on the topics. Feedback and answers to open questions are given using the Large Language Model (LLM) llama3.2-3B with Retrieval Augmented Generation (RAG). The background knowledge contains encyclopedias by the Bundeszentrale für politische Bildung (bpb) and public broadcast video-transcripts. We conducted a preliminary mixed-methods pilot with 4 German speaking participants (age 18 - 45) with 3 of them having a migration background and one of them having completed the citizenship test already. The participants rated the usability as 4.4 ± 0.6 and the perceived improvement over learning without the game as 4.9 ± 0.3 on a 5-point Likert scale showing overall high perceived satisfaction with the game. Three of the participants rated the LLM-generated feedback as helpful, one noted it lacked depth. In qualitative feedback, the participants liked the simple yet effective gameplay loop and proposed additional features such as quests that integrate the learning content in the interaction with NPCs. These results cautiously indicate that an LLM-augmented serious game can deliver engaging, individualized preparation for the German citizenship test. In future work we want to evaluate the system in a trial with a cohort that is preparing for the citizenship test and explore multilingual extensions.

Keywords: Intelligent tutoring game, Intelligent tutoring system, Serious game, Large language models, Artificial intelligence, Leben in Deutschland

1. Introduction

The test “Leben in Deutschland” (English: “Living in Germany”) is the final examination following the orientation course in Germany and is also used as a citizenship test (Federal Office for Migration and Refugees, 2026). The orientation course is part of the integration courses organized by the German Federal Office for Migration and Refugees (BAMF), primarily attended by refugees and newly arrived immigrants. Passing the test is generally required for permanent residence or naturalization; however, individuals who have completed their education in Germany are typically exempt from attending the orientation course. The course covers topics related to the major subject areas of “politics in a democracy”, “history and responsibility”, and “people and society”. These subject areas include many individual topics such as the German legal system, history, and culture. Furthermore, the course discusses rights and obligations in Germany, ways of living together in society, and values — particularly those that are important in Germany, such as freedom of religion, tolerance, and gender equality.

The test itself is a multiple-choice questionnaire with 33 questions with 3 questions that depend on the state where an immigrant lives. To succeed, the applicant must have at least 17 correct answers. The official question database for each state contains a total of 310 questions covering the topics politics, history, the German election system, human rights and discrimination types. 10 of those questions focus specifically on information relevant to a state, such as specifics about the state elections. BAMF publishes the full question database for the test as PDF files and a simple website¹ for study purposes, while there are several other websites that have a

¹ <https://oet.bamf.de/ords/oetut/f?p=534:1:0>

better structured presentation of the questions and therefore allow a better preparation for the test². Before taking the test, most participants will take part in an integration course consisting of up to 700 lessons, with 600 lessons on learning German and 100 lessons orientation course which covers the same topics as the “Leben in Deutschland” test. The test is to be taken after finishing this orientation course. Naturalization itself has several other preliminaries such as living in Germany for at least 5 years or sufficient fluency in the German Language (B1 Level).

Learning all the 310 questions of the database by heart can be very frustrating and a long process, which suggests that motivational support may be beneficial. Serious Games are especially helpful in motivating learners to continue learning. Additionally, the question database does not contain any explanations, when a question is answered incorrectly, so the learners must simply accept the answer labelled as correct or search manually in the course material for the why the answer is correct. Tailored feedback given by human tutors shows the best results in understanding, but the number of skilled tutors is limited and cannot be reached around the clock. Learners may also be too embarrassed to ask for feedback on simple questions. Intelligent Tutoring Systems provide this tailored feedback on a similar level but often struggle with complex world knowledge which is covered by the citizenship test. The recent advances in Natural Language Processing and especially Large Language Models provide a good foundation to build an Intelligent Tutoring System in this domain.

We therefore propose a combination of an Intelligent Tutoring System and a Serious Game – in the following called “Intelligent Tutoring Game” – that could be able to support both higher motivation and tailored feedback and result in a more thorough understanding of the topics covered in the citizenship test. The purpose of the Intelligent Tutoring Game is not to replace the orientation course but stand between that course and the citizenship test in the phase of preparation and learning for the test.

This leads to the Research Questions which we want to answer about our proposed Intelligent Tutoring Game:

RQ1: How can tailored feedback through LLMs be integrated into a Serious Game?

RQ2: To which extent can the integration of tailored feedback by LLMs support Learners in a Serious Game?

2. Related Work

2.1 Serious Games

Serious Games are defined as “games that do not have entertainment, enjoyment or fun as their primary purpose” (Michael and Chen, 2005). They are specifically useful in educational contexts, to motivate learners through fun and interesting interaction. Digital Serious Games additionally have the advantage of being able to adapt to the situation of the learner and being able to use them anywhere and anytime (Gloria, Bellotti and Berta, 2014). Two essential requirements for a good gameplay loop are a clear goal and feedback about the current progress towards the goal (Broin, 2011).

Using a serious game to learn Multiple Choice Questions has not been studied much, especially since the purpose of most serious games lies in the implicit knowledge transfer or learning by doing (Checa and Bustillo, 2020; Tan and Nurul-Asna, 2023; Facchino *et al.*, 2025). In this case, however, it is helpful to prepare for the test directly with the multiple-choice questions from the question database since the final test will use a subset of these questions. A comparison of Multiple-Choice Questions embedded in a Serious Game with Online Quizzes (Karavidas, Skraparli and Tsiatsos, 2024) showed no significant change in performance or enjoyment, but a higher engagement with the content. Integrating the Serious Game with an Intelligent Tutoring system should enable a better understanding of the questions and higher engagement while learning.

2.2 Intelligent Tutoring Systems

Intelligent Tutoring Systems (Graesser, Conley and Olney, 2012) are digital learning systems that track the psychological state of the learner, represent the topic to be understood and give tailored feedback with the aim to support the learner in gaining a thorough understanding of the topic, founded on pedagogical principle. In comparison to serious games, they focus more on learning the topic, while serious games add an engaging environment around the task. A combination of an Intelligent Tutoring System and Serious Games shows that they increase motivation for learning while having no impact on the learning effectiveness in the long term (Jackson and McNamara, 2013).

² See e.g. <https://www.lebenindeutschland.eu/>, <https://einbuengerungstest-fragen.de/> or <https://lebenindeutschland-test.de/>

2.3 Large Language Models and Retrieval Augmented Generation

Large Language Models (LLMs) have been a constant topic of debate and additional advances since the release of ChatGPT in 2021. LLMs generate text based on a given starting text, called prompt. The basic building block behind it is the Transformer Architecture (Vaswani *et al.*, 2017), which uses the Attention Mechanism to connect relevant information in the prompt using the parameters that were learned while training on a large amount of text data from the internet and afterwards tuned to respond to instructions instead of just continuing text. These models are inherently not able to always give correct answers, and the correctness depends a lot on the prompt given (Sclar *et al.*, 2024). Additionally, some information might not have been in the training data, so the LLM might not be able to answer correctly. To improve the accuracy of answers on specific topics, such as the German Citizenship Test, a common approach is Retrieval Augmented Generation (RAG), where the most relevant content for a question to the LLM is inserted into the prompt (Lewis *et al.*, 2020). This reformulates the task for the LLM from a complete question answering to a content rephrasing problem if the retrieved content is answering the question. Retrieving the content can be done using information retrieval approaches (Hambard and Proença, 2023), such as term matching, embedding models that map the text to a vector representation that more closely resembles the semantic meaning of the text, or a combination of both. Additional RAG paradigms improve the accuracy using e.g. query rewriting before or reranking after the retrieval step (Gao *et al.*, 2024). Al-Abri (2025) has shown that LLMs can be successfully employed in an Intelligent Tutoring System, but students need to be aware of their varying reliability, and Stamper, Xiao and Hou (2024) advocate for a more thorough evaluation of the impact of LLMs, while also considering theoretically grounded methods for feedback generation.

2.4 Serious Games for Citizenship Tests

To our knowledge, there is only one other Serious Game that covers the Citizenship test called “Leben in Deutschland – spielend integrieren” (engl. “Living in Germany – playful integration”).³ This is a board game first developed in 2016 for up to 6 players. Those players travel on a Germany map, and based on the color of the new location, they must draw a card – which is one of the multiple-choice questions of the citizenship test. The focus of the game is to get into discussions on the questions in the group and thereby implicitly gaining knowledge about Germany.

Another board game is “So you want to be an American”⁴, which covers the American citizenship test. This board game covers the possible ways an immigrant can take, to become an American citizen. Its focus lies on making American citizens aware of and reflect on the complexities and sometimes even arbitrariness of the regulations on Immigration in America.

3. Design

We introduce a design of an Intelligent Tutoring Game for learners of multiple-choice questions. On the one hand, this includes the design of the Serious Game, which has the purpose of motivating the learner while also not distracting them too much from the learning objective. On the other hand, the Intelligent Tutoring System has the purpose of providing a tailored learning experience for the learner and enabling a more thorough understanding of the topics.

3.1 Design of the Serious Game

The Intelligent Tutoring Game situates the Learners in a familiar and interesting but also simple 2D environment that matches the content to be learned, while still looking playful. In this case the test material is from the German Citizenship Test, so the learner is situated in a German town. The player starts in front of their own house and can move freely through the town. Several buildings can be entered and have a specific purpose in the progression of the game and learning objectives.

First, the player is prompted to enter the test center, which corresponds to the final testing environment in real life. Instead of going there only once, this is the place where the player can practice multiple-choice questions. Each time the player starts a new practice test, they get a set of 33 system-selected questions. The selection is discussed in section 3.2. After a test is finished, they can go to a separate room in the test center where the

³ <https://www.lid-integration.de/>

⁴ <https://www.soyouwanttobeanamerican.com/>

results of all practice tests can be viewed. The player can choose to view all questions that have been practiced until this point, the questions from the last practice test or only the incorrectly answered questions. Afterwards, they can now choose to continue practicing, receive tailored feedback on the incorrect questions or ask open questions.

To receive tailored feedback, the player can go to another part of town, where three module buildings correspond to the major subject areas “politics in a democracy”, “history and responsibility”, and “people and society”. Each of those buildings is structured in the same way with a help desk where the learners receive an overview of how many questions they have gotten right in the respective subject and specific LLM-Generated Feedback on each incorrectly answered question, covering why the selected answer is incorrect and why the correct answer is correct.

To ask open questions, the player can go to the library and talk with the librarian. The librarian is an LLM-based assistant which has specific background information on the test material. The learners can therefore receive answers to their individual questions regarding specific topics. In this way, the library complements the module buildings. This supports the learning objectives according to the Revised Bloom Taxonomy (Airasian *et al.*, 2001), particularly understanding and analyzing. As a result, the learning content is also reviewed and reinforced at the individual learner level. Learners receive concrete answers to questions that particularly interest them and that may have been explained more generally in the module buildings.

Several small aspects add to playfulness and motivate the learner to continue. Small successes are rewarded verbally, e.g. finding the way to each building for the first time or after finishing each test. The progress of how many questions have been answered correctly is always visible in the bottom right of the screen, to give the players a sense of accomplishment. It is, however, not too prominent, to not feel like an impossible goal.

3.2 Design of the Intelligent Tutoring System

The basis of an Intelligent Tutoring System is a representation of the domain knowledge. This is realized in this system in two parts, first the multiple-choice questions with the knowledge about the correct answers and second the additional knowledge about those respective topics, which is built from a collection of trustworthy website content and video or podcast transcripts. The student module of the Intelligent Tutoring System represents an assessment of the current knowledge of the student about the domain. This is realized by storing the results of each practice test. Since a single correct/incorrect response can be made by mistake, the mastery of a specific question is achieved when the last 2 attempts at the question were successful.

The pedagogical module of the Intelligent Tutoring System is built up of several aspects. First, the questions for a practice test are selected depending on the current estimated knowledge of the student. If the student has not practiced each question at least once, 22 of the questions are selected randomly from questions that have not been practiced. The other questions are selected randomly from the already practiced questions, where the probability of selection of a Question is higher if the number of failed attempts is higher. Mastered questions are excluded unless there are no other questions left, which would only be the case when all questions have been trained. Second, the students receive tailored feedback on failed questions in the module buildings introduced in section 3.1. The tailored feedback is built by providing the original question, the learner answer and the correct answer to a LLM with one additional relevant document snippet retrieved from the domain knowledge. The feedback aims to foster learners understanding of the difference between their answer and the correct answer. And third, open questions can be asked to the librarian, also introduced in section 3.1. The answers are also based on the most relevant document snippet retrieved from the domain knowledge, with an additional persona that is given to the librarian.



Figure 1: The town with module buildings (left), test center and library (right)

4. Implementation

The game is implemented as a 2D tile-based top-down RPG using PyGame as engine, which was selected due to easy prototyping and integration with LLM-Tooling. The tileset by LimeZu (2021, 2025) was utilized for the design of the town and the characters shown in Figure 1. The implementation of the game is released publicly under the MIT License.⁵

The integration with LLMs is done locally using Ollama⁶ and the vector-embedding database for RAG is set up using ChromaDB⁷. The LLM model llama3.2:3b⁸ developed and trained by Meta was used for generating answers and nomic-embed-text⁹ developed by Nomic for embedding in background knowledge. This selection was performed based on the size of the language models, open weights and best benchmark performance at the time of first development in 2025.

The background knowledge contains 328 lexicon entries from the German Federal Agency for Civic Education (bpb), specifically from the „Handbuchwörterbuch Politisches System“ (Andersen *et al.*, 2021) and the „einfach POLITIK: Lexikon“ (Meyer, Vock and Schüller-Ruhl, 2022), 9 articles from bpb, Goethe-Institut, and the State Agency for Civic Education of Baden-Württemberg, as well as 3 podcast transcripts and 33 video transcripts from German public broadcasting channels. This amounts to a total of 758 179 words, embedded in chunks of at most 1000 tokens to not break up individual topics, resulting in 5062 chunks.

We used the naïve RAG approach (Gao *et al.*, 2024) to retrieve the most similar chunk from the vector database. During testing, inserting only the most relevant chunk yielded shorter and more concise results than inserting several similar documents.

The LLM-prompt for the librarian is very simple, with a short persona description, the requirement to answer in a maximum of 7 sentences, the most relevant chunk for the user question and then the user question. The prompt template for tailored feedback on a failed question is shown in Figure 2.

Using this information: **{chunk}**
 Clarify the following multiple-choice question with an incorrect answer in maximum 100 words
 Question: **{question}**
 Options: A. **{option_a}** B. **{option_b}** C. **{option_c}**
 D. **{option_d}**
 The correct answer is option '**{correct_option}**'.
 Why is this the correct answer?
 The user selected option: '**{selected_option}**'.
 Why is this answer incorrect?

Figure 2: Prompt for Tailored Feedback on failed questions. Template is originally German and was translated for accessibility

⁵ <https://gitlab.rz.uni-bamberg.de/cogsys/its/leben-in-deutschland/>

⁶ <https://ollama.com/>

⁷ <https://github.com/chroma-core/chroma>

⁸ <https://ollama.com/library/llama3.2>

⁹ <https://ollama.com/library/nomic-embed-text>

Table 1: Questionnaire for the evaluation of the game

Index	Item	Question Type	Question (translated from German)
1	i1: Usability	5-point Likert	The controls were intuitive.
2		5-point Likert	I could easily orient myself in the game.
3		5-point Likert	The test procedure in the test center was practical and clear.
4		Open Question	What did you like most about the game?
5		Open Question	What did you find confusing or difficult?
6	i2: Design	5-point Likert	The design of the game is good.
7		5-point Likert	The content was presented in an understandable way.
8		5-point Likert	Being able to choose between "All questions" / "Last questions" and "only incorrect" in the test result analysis makes sense.
9		5-point Likert	The game was entertaining
10		Open Question	Did you feel that the learning was well integrated into the game's storyline?
11		Open Question	How did you find the gameplay?
12		Open Question	Do you have any suggestions on how the game could be improved?
13	i3: Learning effectiveness	5-point Likert	I have learned something new in the game
14		5-point Likert	The game helps to memorize the content
15		Open Question	Do you think the game makes learning clearer?
16		Open Question	Do you think the game makes learning easier?
17	i4: Serious Game vs without	5-point Likert	It is easier to learn in a game with buildings than with a PDF-File.
18		5-point Likert	The game with different locations makes learning easier.
19		5-point Likert	The test result analysis in the small room in the test center makes sense.
20		5-point Likert	The library promotes learning, especially by being able to freely formulate own questions for the teacher.
21		5-point Likert	The different locations have motivated me to continue playing.
22		Open Question	What are the advantages of this game over a PDF file with questions?
23		Open Question	Did you have more fun learning this way than with a PDF file full of questions?
24	i5: Recommendation	5-point Likert	I would play the game again.
25		5-point Likert	I would recommend the game.
26		Open Question	What was your overall impression of the game?
27	i6: LLM-Performance	5-point Likert	The explanations in the module buildings are helpful.
28		Multiple Choice	The teacher's answers in the library are: well-explained, informative, misleading, insufficient.
29		Multiple Choice	The explanations in the module buildings are: clear, incorrect, useful and helpful, too brief, too long.

5. Evaluation

Abdellatif, McCollum and McMullan (2018) state that to measure the quality of a Serious Game on learning, several aspects must be evaluated. This includes the usability of the system itself, the engagement of the users with the game, the impact of the game on the motivation to learn, the understandability of the information given by the game and the overall user experience.

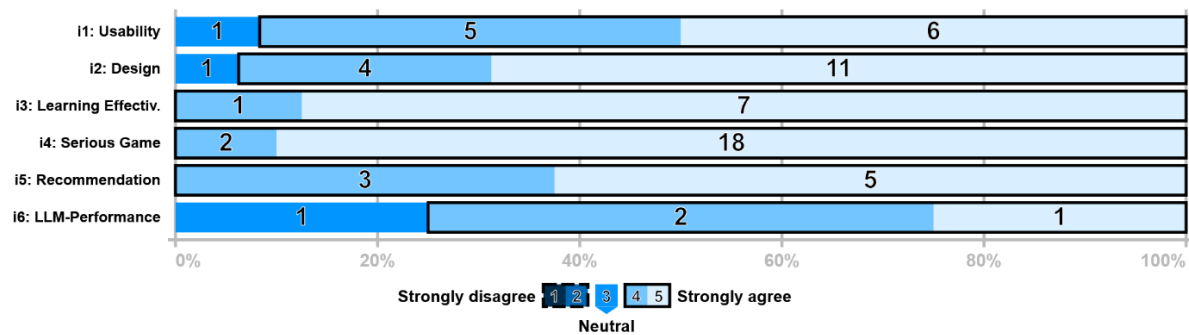


Figure 3: Results of the Likert-Scale Questions grouped by item

5.1 Mixed-methods Pilot

As a preliminary evaluation of the presented game, we performed a mixed-methods pilot. The purpose of the pilot is not to provide a complete evaluation as presented by Abdellatif, McCollum and McMullan (2018), but instead give first directions for further improvements before performing a complete evaluation. It started with the participants experiencing one full cycle of the game play loop, which included going to the test center, taking a test, looking at the results analysis, going to the buildings for each module to get feedback and lastly going to the library for open questions. Afterwards, the participants were given a questionnaire, starting with demographic questions and knowledge about the citizenship test and video games. The rest of the questionnaire consists of a total of 29 questions in 6 different categories, presented in Table 1. 5 questions on the usability of the Serious Game, 7 questions on the visual and conceptual design of the game, 4 questions on the general learning effectiveness, 7 questions on the comparison of learning with the Serious Game versus without the Game, 3 questions on recommendation of the game and 3 questions on the Performance of the LLM-generated answers. 17 of the questions are in the form of a 5-point Likert scale ranging from “strongly disagree” to “strongly agree”, 2 questions are Multiple choice and 10 are open questions.

The pilot had 4 participants. P2, P3 and P4 were between 18 and 27 years old, P1 was 45 years old. P2, P3 and P4 had mid-to-high experience with video games, P1 had none. P1, P2 and P4 have a migration background and P1 has passed the citizenship test before, P3 and P4 had previous knowledge on the topic. P3 and P4 study information systems.

The results of the Likert-Scale items presented in Figure 3 show in total high satisfaction with the game by the participants. On average, the usability is rated as 4.4 ± 0.6 , the design as 4.6 ± 0.6 , the learning effectiveness as 4.8 ± 0.3 , the perceived improvement over learning without the game as 4.9 ± 0.3 , the LLM-Performance as 4 ± 0.7 and the recommendation as 4.6 ± 0.5 . Three of the participants rated the LLM-Performance in the multiple-choice questions as well-explained, informative, clear as well as useful and helpful, while one of the participants selected it was insufficient and too brief.

The responses to the open questions are accumulated as qualitative results. The participants asked for more guidance in the game, e.g. on the controls, next steps and the current progress. They also asked for the ability to quit a test and to switch back and forth during the test. Additionally, they had some ideas for extending the game with more motivational feedback (“Nice!”, “Keep it up!”), conversations with other NPCs possibly with indirect explanations of misunderstood topics, an info button on a question during a test, and more broadly in-game events, other locations or collectibles. Positive feedback included satisfaction with the test result analysis and the ability to ask open questions in the library and the participants liked the simple yet effective design of the game.

Using this data: [Section in the Video Transcript of „Ist die Meinungsfreiheit bei uns in Gefahr?“ by MrWissen2Go on the topic of criminal complaints and Free Speech]
Answer the following question in maximum 7 sentences: Why should I follow house rules?
Answer: It's important that you follow the house rules because you're a guest in a public space. This means adhering to certain rules to respect others and ensure that everyone enjoys a pleasant atmosphere. If you don't follow the rules, it can lead to discomfort or conflict. You should also follow the rules because it can help you control your own feelings and opinions and avoid being overly influenced by others. It's also important to follow the rules because it shows respect for others. If you don't follow the rules, there may be consequences. It's always better to follow the rules to create a positive atmosphere.

Figure 4: Filled in Prompt and Generation Result for an open question on the Topic of “Why should I follow the house rules”. The retrieved Background Information does not match the question directly, but the result generated gives mostly good advice and explanation on the topic. Prompt and result are translated from German

Using this information: [Section in the Entry “Culture Politics” of the „Handbuchwörterbuch Politisches System“ (Andersen et al., 2021) on the Topic of the “German Cities Day”]
Clarify the following multiple-choice question with an incorrect answer in maximum 100 words:
Question: Which city was divided into four sectors after World War II?
Options: A. Frankfurt B. Berlin C. Hamburg D. Munich
The correct answer is option 'Berlin' – Why is this the correct answer?
The user selected option: 'Frankfurt' – Why is this answer incorrect?
Answer: The correct answer is option B: Berlin. The incorrect answer is option A: Frankfurt. Frankfurt was divided into four sectors after World War II, but Berlin was not. Berlin was, in fact, divided into four sectors controlled by the Allies (the United States, the United Kingdom, France, and the Soviet Union). Frankfurt, on the other hand, was not divided into sectors.

Figure 5: Filled in Prompt and result for an incorrect answered question. The retrieved Background information is not relevant for the answer, and the answer generated is mostly generic and not helpful. Prompt and result are translated from German

5.2 LLM-response Quality

The LLM-generated results are sometimes helpful and at other times very broad. Though this was not evaluated in detail yet, we want to show two situations of each of the use cases where the LLM was integrated. Figure 5 shows an example for an open question, where the result looks mostly helpful for a learner, despite having the RAG-System not finding a relevant entry. Figure 5 shows an example for the multiple-choice explanation. While the result matches the question, it is very generic and would not enhance the learners' understanding of that question. This kind of behavior happens several times, but the quantity was not evaluated yet. Sometimes the LLM also refuses to explain questions about Germans Nazi-History. but misleading answers were not noticed yet though it cannot be ruled out that the LLM generates them.

6. Conclusion, Limitations and Future Work

The Intelligent Tutoring Game presented in this paper is a promising starting point for creating an engaging learning experience for the German Citizenship Test. With the ability to receive tailored feedback with the support of LLMs, it creates a supportive environment to prepare for the test.

As described in Section 4, the presented game was implemented using Python but was since then converted to Godot for easier publishing. The conversion was done with the assistance of Agentic AI Systems. Additionally, since the initial development of the game, the official question database was updated and is no longer structured into the broad categories “politics in a democracy”, “history”, and responsibility” and “human and society” but rather more fine grained into the topics “politics”, “history”, “german election system”, “human rights”, “discrimination types”, and 10 state dependent questions. This update requires a new structure for the feedback options and limits the longevity of the game, if those questions are regularly changed. Some of the feedback and enhancement suggestions by the participants have already been integrated.

The limitations of the mixed methods pilot are manifold. The group size is very small, which limits the validity of the results. The questionnaire was not based on validated measures which limit the comparability with other systems. The participants had only a very short amount of time to play the game, so the long-term effects of

learning with the game and potential frustrations and effects on motivation are not evaluated. The questions on learning effectiveness are only based on subjective statements and may not correlate with the actual learning effectiveness. Some of the questions are also suggestive, making the participants lean towards a positive answer.

Cautiously we can answer our research questions:

RQ1: How can tailored feedback through LLMs be integrated into a Serious Game? The integration can be done by giving the learners feedback in a contextualized setting, giving specific feedback on course content in an evaluation context and answering open questions with a virtual assistant embedded in the game.

RQ2: To which extent can the integration of tailored feedback by LLMs support Learners in a Serious Game? The self-reported high learning effectiveness in our evaluation shows promising results while the feedback was sometimes not detailed enough. However, the evaluation is very limited as discussed above.

The following design lessons should be considered in a similar endeavor: comprehensive background knowledge is very important, and retrieval must be evaluated extensively. Regarding LLM-integration, we found that it brings the most value where personalized feedback is necessary and cannot be created by traditional means, such as in the case of open questions. The multiple-choice explanations on the other hand could be written manually or with LLM-Assistance, since they are very similar for all learners. And lastly the very simple game design might not be engaging enough in the long term.

In future work we want to update the Godot-Based game with the new question database and add more NPCs and objects to interact with, which allows for a more natural exploration of the topics. The simplicity of the game should however be kept intact, not to distract the learners from the original objective but balancing motivation and progress. The game is to be evaluated with a larger cohort that is preparing for the Citizenship Test and evaluating the learning effectiveness explicitly. An additional extension to the game would be to introduce the game early in the orientation course and integrate all the information discussed as explorational material.

Ethics Declaration: No ethics declaration was required for this research.

AI Declaration: DeepL was partially used for translation from German to English. Consensus was partially used to find references in the Related Work section. All references have been checked for appropriateness and relevance. Generative AI was not used in writing the paper.

References

- Abdellatif, A.J., McCollum, B. and McMullan, P. (2018) "Serious games: Quality characteristics evaluation framework and case study," *2018 IEEE Integrated STEM Education Conference (ISEC). 2018 IEEE Integrated STEM Education Conference (ISEC)*, Princeton, NJ, USA: IEEE, pp. 112–119. Available at: <https://doi.org/10.1109/ISECon.2018.8340460>.
- Airasian, P. et al. (2001) *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.
- Al-Abri, A. (2025) "Exploring ChatGPT as a virtual tutor: A multi-dimensional analysis of large language models in academic support," *Education and Information Technologies*, 30(12), pp. 17447–17482. Available at: <https://doi.org/10.1007/s10639-025-13484-x>.
- Andersen, U. et al. (eds.) (2021) *Handwörterbuch des politischen Systems der Bundesrepublik Deutschland*. Wiesbaden: Springer Fachmedien. Available at: <https://doi.org/10.1007/978-3-658-23666-3>.
- Broin, D.Ó. (2011) "Using a Criteria-Based User Model for Facilitating Flow in Serious Games," *2011 Third International Conference on Games and Virtual Worlds for Serious Applications. 2011 Third International Conference on Games and Virtual Worlds for Serious Applications*, pp. 63–69. Available at: <https://doi.org/10.1109/VS-GAMES.2011.15>.
- Checa, D. and Bustillo, A. (2020) "A review of immersive virtual reality serious games to enhance learning and training," *Multimedia Tools and Applications*, 79(9), pp. 5501–5527. Available at: <https://doi.org/10.1007/s11042-019-08348-9>.
- Facchino, A.P. et al. (2025) "The use of serious games for psychological education and training: a systematic review," *Frontiers in Education*, 10. Available at: <https://doi.org/10.3389/feduc.2025.1511729>.
- Federal Office for Migration and Refugees (2026) *Integration courses: The content and stages of the procedure, BAMF - Federal Office for Migration and Refugees*. Available at: <https://www.BAMF.de/EN/Themen/Integration/ZugewanderteTeilnehmende/Integrationskurse/InhaltAblauf/inhaltAblauf-node.html> (Accessed: April 29, 2026).
- Gao, Y. et al. (2024) "Retrieval-Augmented Generation for Large Language Models: A Survey." arXiv. Available at: <https://doi.org/10.48550/arXiv.2312.10997>.
- Gloria, A.D. et al. (2014) "Serious Games for education and training," *International Journal of Serious Games*, 1(1). Available at: <https://doi.org/10.17083/ijsg.v1i1.11>.

- Graesser, A.C., Conley, M.W. and Olney, A. (2012) "Intelligent tutoring systems," *APA educational psychology handbook, Vol 3: Application to learning and teaching*. Washington, DC, US: American Psychological Association (APA handbooks in psychology®), pp. 451–473. Available at: <https://doi.org/10.1037/13275-018>.
- Hambarde, K.A. and Proença, H. (2023) "Information Retrieval: Recent Advances and Beyond," *IEEE Access*, 11, pp. 76581–76604. Available at: <https://doi.org/10.1109/ACCESS.2023.3295776>.
- Jackson, G.T. and McNamara, D.S. (2013) "Motivation and performance in a game-based intelligent tutoring system," *Journal of Educational Psychology*, 105(4), pp. 1036–1049. Available at: <https://doi.org/10.1037/a0032580>.
- Karavidas, L., Skraparli, G. and Tsiatsos, T. (2024) "A Comparison between Online Quizzes and Serious Games: The Case of Friend Me," *Computers*, 13(3), p. 58. Available at: <https://doi.org/10.3390/computers13030058>.
- Lewis, P. et al. (2020) "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 9459–9474. Available at: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html> (Accessed: April 29, 2026).
- LimeZu (2025) *Modern Exteriors* [Tileset]. Available at: <https://limezu.itch.io/modernexteriors> (Accessed: April 23, 2026).
- Meyer, D., Vock, R. and Schüller-Ruhl, T. (2022) *einfach POLITIK: Lexikon*. Edited by W. Hilpert. Bundeszentrale für politische Bildung/bpb. Available at: <https://www.bpb.de/kurz-knapp/lexika/lexikon-in-einfacher-sprache/> (Accessed: April 29, 2026).
- Michael, D.R. and Chen, S.L. (2005) *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade.
- Sclar, M. et al. (2024) "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting," *The Twelfth International Conference on Learning Representations*. Available at: <https://openreview.net/forum?id=RIu5lyNXjT> (Accessed: April 29, 2026).
- Stamper, J., Xiao, R. and Hou, X. (2024) "Enhancing LLM-Based Feedback: Insights from Intelligent Tutoring Systems and the Learning Sciences," in A.M. Olney et al. (eds.) *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*. Cham: Springer Nature Switzerland, pp. 32–43. Available at: https://doi.org/10.1007/978-3-031-64315-6_3.
- Tan, C.K.W. and Nurul-Asna, H. (2023) "Serious games for environmental education," *Integrative Conservation*, 2(1), pp. 19–42. Available at: <https://doi.org/10.1002/inc3.18>.
- Vaswani, A. et al. (2017) "Attention is all you need," *Advances in Neural Information Processing Systems. 31st Conference on Neural Information Processing System*, Long Beach, CA, USA. Available at: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html> (Accessed: September 30, 2024).