EVALUATING TRUSTWORTHINESS, USABILITY AND EXPLAINABILITY OF AN EDUCATIONAL PATHWAY RECOMMENDATION SYSTEM THAT USES A LARGE LANGUAGE MODEL

T. Zylowski[1,2], N. Sautchuk Patricio[2], W. Hettmann[2], K. Anderer[2,3], M. Wölfel[1,2], P.A. Henning[2]

[1]University of Hohenheim (GERMANY)
[2]Karlsruhe University of Applied Sciences (GERMANY)
[3]Karlsruhe Institute of Technology (GERMANY)

Abstract

Advising and supporting pupils in their career choices and the associated further education paths is becoming increasingly important. Current studies show that there is a high drop-out rate in bachelor's degree courses because the ideas and expectations of the pupils and the content offered in the course do not overlap sufficiently. While pupils widely started to use ChatGPT as a tool for educational purposes, it is not specifically designed to provide career recommendations, lacking domain knowledge and country-specific requirements. Another drawback is that there is no guidance in the conversation and a conversation is driven by user questions, not by the system itself.

To meet this challenge, we have developed an AI-based assistant to provide individual advice on career choices and identify suitable educational paths. The assistant is a chatbot based on a modern open-source large language model (LLM) hosted in a data-sovereign manner and adapted to the educational domain using prompt engineering. This allows for an open conversation that is more human-like. In a conversation, the goals, ideas, experiences, and skills of the pupils are determined and a user profile is created. This user profile is used to make several recommendations for an occupation. If the user is not satisfied, the conversation is deepened further for more details. As soon as an occupation matches the pupil's requirements, several possible educational paths are calculated. These can consist of entry into the German 'dual' system of professional education, or different bachelor and master courses leading to the selected occupation and adapted to the pupil's profile.

To evaluate the newly developed assistant, we designed and conducted an experiment with pupils to evaluate the usability of the assistant in recommending careers and educational pathways. We created two additional assistants: a *form-based assistant* that functions as the baseline of the comparison and an *intent-based chatbot* that maps a user utterance to an intent and triggers an action. That way the conversation is prescribed with a fixed set of utterances and the assistant is not able to have open conversations. A benefit of the intent-based chatbot is that the conversation is very structured and guided by the assistant.

An important aspect of such systems is whether they are trusted and accepted by the users. Therefore, we examine in the user test the effect of the assistants on trustworthiness and associated conditioning factors. These include perceived competence, autonomy, and anthropomorphization of the assistant. We investigate the perceived ability of the assistants to explain their utterances and recommendations.

With this paper, we contribute to the development of a trustworthy educational path recommendation that can give personalized career suggestions. Therefore, we compare three different assistants evaluating how they make recommendations that pupils feel they can trust.

Keywords: Generative Artificial Intelligence, User Study, Conversational User Interface, Chatbot, Educational Pathway Recommendation, Large Language Model.

1 INTRODUCTION

Choosing a career and educational path represents a significant decision for pupils, influenced by various factors such as societal and peer-mediated expectations, available opportunities, and personal aspirations. The degree of satisfaction and genuine enjoyment derived from the chosen academic program, coupled with a strong interest in the course, significantly correlates with academic achievement and successful completion of undergraduate studies [1]. Additionally, a sense of belonging to the

ISBN: 978-84-09-62938-1

university environment is linked with enhanced well-being, increased academic motivation, and reduced likelihood of dropping out [2]. Nonetheless, many pupils face challenges in discerning the most suitable educational and career paths, as well as educational institutions aligning with their values, due to their career aspirations and abilities often remaining latent or unrecognized [3]. Further, pupils often lack good career education and guidance in school [4]. Consequently, it is critical to identify and assess pupils' career aspirations and skills early on and to empower them to develop a clear understanding of their potential career paths.

Various assessment methods are available to evaluate pupils' skills and requirements for career planning. These methods range from traditional paper-based questionnaires manually evaluated to digital versions analyzed either manually, semi-automatically, or automatically. However, for many involved parties, the use of these questionnaires remains opaque, making it challenging to understand how the suggested educational pathways align with their responses. Furthermore, digital questionnaires often have limited input capabilities, as they typically do not analyze the content of text fields due to the necessity for advanced natural language processing (NLP) techniques. This constraint inhibits personal expression, which is vital for effectively addressing pupils' career aspirations. To address this issue, it is essential to rethink career planning methods, integrating innovative approaches to recommend career paths that closely align with each pupil's strengths and goals.

Chatbots present a promising solution for evaluating individual career plans expressed in natural language. They possess the capability to comprehend and engage in natural language, facilitating deeper conversations and a more profound understanding of pupils' career expectations. Within chatbot interactions, structured survey methods such as Likert scales or multiple-choice questions can be also incorporated. Intent-based chatbots use machine learning to classify user inputs and deliver predefined responses, offering structured dialogues albeit with limited flexibility. Conversely, chatbots powered by large language model (LLM) chatbots generate responses directly from input, utilizing a language model that has been trained on a vast amount of data. This allows for error-tolerant, open-ended conversations and enables customization to individual user preferences through prompt engineering and fine-tuning [5]. It is also possible to change the character of an LLM using suitable prompts and thus adapt it to user preferences [6]. However, LLM-based systems may occasionally generate incorrect information that is frequently referred to as *hallucination* [7]. LLM-based chatbots are perceived as more trustworthy compared to intent-based chatbots because they offer plausible-sounding answers to the user [8]. However, the fact that they sometimes produce hallucinations might lead in the long term to reduced trust or dissatisfaction in the context of career counseling with LLM-based chatbots. They might also lack guidance because they are generally designed to give answers to the user, not to first collect relevant data to explore different study or occupation directions and provide career recommendations. While LLM-based chatbots often depend on user guidance, in advisory contexts, the chatbot needs to take the lead in steering the conversation. ChatGPT [9], which is one of the most commonly used LLM-based chatbots nowadays, is used by many high school and college pupils as a tutor, according to a study conducted by Intelligent [10].

To address the complexities of career planning, we have developed an AI-based assistant aimed at providing personalized guidance on career choices and recommending suitable educational paths. This assistant operates as a chatbot, employing an open-source LLM hosted within a data-sovereign environment as part of the MERLOT initiative (MarkEtplace foR LifelOng educaTional dataspaces and smart service provisioning) [11]. MERLOT establishes a secure educational data space, fostering data exchange among participants to deliver cutting-edge educational services and applications to end-users through a marketplace. We compare the AI-based assistant to two other assistants, one that is *intent-based* and uses predefined structures for predefined intents of the user that are recognized by natural language processing [12], and a *form-based* assistant that serves as a baseline. While more flexibility and higher user's trust is anticipated for the LLM-based assistant, more guidance is anticipated for the intent-based system.

In practical scenarios of educational pathway recommendations, the acceptance of chatbots relies, for instance, on their perceived benefits, ease of use, and ability to provide trustworthy guidance, correctness, and recommendations. Additionally, there is a presumed connection between explainability and trust. The effectiveness of a chatbot depends on the quality of dialogue, yet it remains uncertain which characteristics are most important for users in career recommendation contexts. This paper outlines the experimental design and the lessons learned from a pre-test study, focusing on interactions between pupils and chatbots with a focus on usability, trustworthiness and explainability.

2 METHODOLOGY

In the following subsections, we delineate our methodology concerning the establishment of the experiment and the execution of measurements utilizing structured questionnaires focused on behavioral observations and self-assessment. The motivation behind a pre-test was multi-folded. First, the pre-test should give a proof of concept and show that the combination of a workshop about GPT and prompts together with the conductance of the user study is engaging for the pupils. The pupils should show interest in the topic and get the feeling that their feedback helps in the development of a career assistant. Further, although the role of attention digital school experiences has been questioned, the pre-test should show whether the length of the questionnaires and the total length of the study is manageable with respect to the attention span of the pupils, which is dependent on age [13]. Lastly, the questionnaires were not specifically designed for pupils and the language level might need to be adjusted. To sum up, a pre-test serves as a general exploration and feedback loop to improve our study concept.

2.1 Procedure of the Pre-Test Study

For the pre-test study, we adopted a quasi-experimental design employing a multi-group plan. Three distinct versions of the assistant were developed for testing: *form-based* (F), *intent-based* (R), and *LLM-based* (L). All participants completed identical questionnaires, one administered before and one after the assistant test. Each participant was randomly assigned to only one of the three assistants with a unique user ID to maintain anonymity and establish a connection between their interactions with the assistants and the questionnaire data. For this purpose, participants were required to input their IDs in the questionnaires and the assistant interface at the beginning. We tested to ascertain the reliability of participants in adhering to this procedure.

The pre-test study was undertaken during a pupil visitation day at Karlsruhe University of Applied Sciences, which focused on career and course orientation. In the afternoon sessions, pupils were given the option to participate in one of two workshops. These workshops, each lasting 2 hours, were organized and conducted by different research institutes and departments within the university.

We hosted a workshop for pupils focused on AI-driven career planning. The session comprised an introduction to ChatGPT followed by discussions, testing of the three developed assistants, and a prompt engineering workshop. Opting for a workshop format aimed not only to enhance participation motivation but also to provide tangible benefits to the pupils beyond mere participation in the study. The participants were in their final year (12th grade) at a secondary school, specifically a “Realschule Plus” in Germany, underscoring the emphasis on career and academic guidance. A total of eight pupils and one teacher engaged in the workshop.

During the workshop, following a brief welcome, participants were initially asked about their current career planning status and their familiarity with artificial intelligence (AI) and ChatGPT. Subsequently, a concise introduction to ChatGPT was provided. Participants were then presented with a slide outlining the testing procedure with the assistants and informed that they could seek clarification by raising their hand at any point. After completing the initial questionnaire, testing of the three assistant versions started in parallel which lasted approximately 15 to 20 minutes. Upon completion, participants were asked to fill out the final questionnaire. In the last part of the two-hour workshop, participants received insights on prompt engineering, followed by an open discussion exploring the societal impacts of AI, wherein participants actively engaged. Finally, participants were invited to provide feedback on both the assistants and the workshop as a whole.

2.2 General Flow of the Assistants

The primary task of the three assistants is to gather information regarding a participant's future career aspirations, skills, and current educational steps to create a user profile. How this profile is constructed varies across the assistants. However, all three assistants utilize the user profile in the same way to generate two occupation recommendations. Participants can either opt for one of these recommendations or indicate that none are suitable by clicking on the corresponding buttons, as can be seen in Figure 1. Further skill-related questions are posed if none of the suggestions align with the participant's preferences. For the form-based and intent-based assistants, these skill questions were predetermined by humans, while for the LLM-based assistant, they were autonomously generated by the LLM itself. The user profile is subsequently updated based on the participant's responses to these skill questions. Using the updated profile, two new occupation recommendations are proposed. Upon selecting a recommended occupation, participants are given the choice to specify a preferred location to study. Following this, at least one educational pathway is computed and presented in the format of a timeline, detailing specific educational steps at specific universities or outlining general training courses available in Germany, as shown in Figure 2. In case of having more than one educational path, they are shown in different tabs.

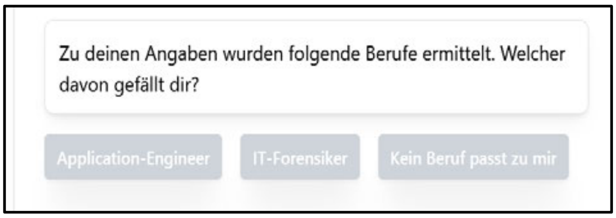

Figure 1. This figure shows the buttons that are displayed after occupations are matched to the user. Two buttons display different occupation recommendations, while a third option is available for the case where no occupation is satisfying to the user.

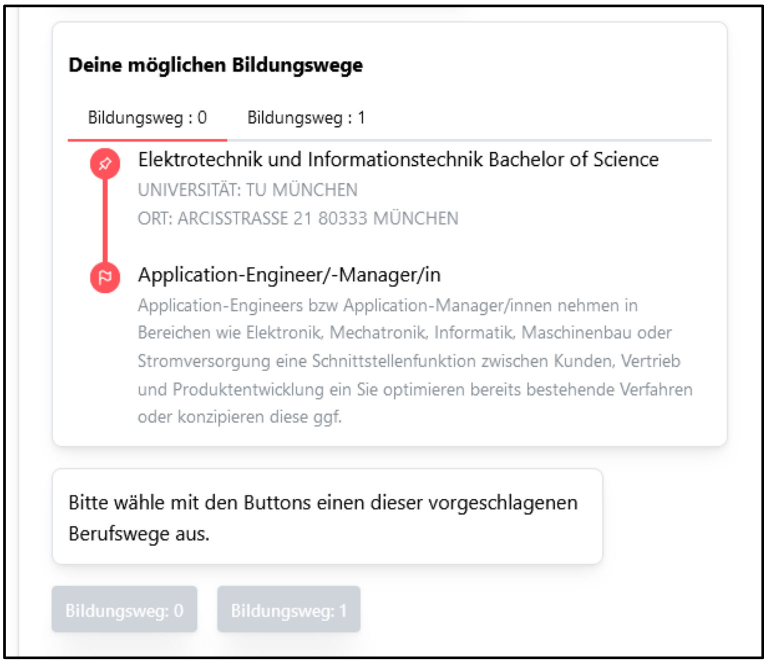

Figure 2. This figure presents the options for educational pathways. The user can click on one of them to display more details about the pathways. Below, the user then has the option to select a pathway.

2.3 Form-based Assistant

Various career advising tools rely on form-based interfaces, such as O*NET Interest Profiler [14] and Check-U [15]. Therefore, for the control group, an assistant in the form format was selected, as displayed in Figure 3. The data collected from the form, constituting the user profile, is sent to the recommendation engine for occupations and educational paths, in a similar process used with the other two assistants. However, compared to the other two assistants, the interface is not conversational. While the questions asked are identical to those in the intent-based chatbot, they are neutrally formulated and devoid of the characteristic perspective of a chatbot entity. Participants are initially asked about their existing career aspirations or preferred educational paths. If they lack a specific direction, they are presented with eight questions aligned with the RIASEC assessment framework, designed to gather insights about pupils' interests, preferred activities, beliefs, abilities, values, and characteristics in order to classify them into six personality types: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional [16]. Examples include questions about their primary subject or topics of interest and their approach to handling new creative ideas.

If the participant deems the initial occupation recommendation to be unsatisfactory, supplementary skill-related inquiries are posed to elucidate their capabilities in greater depth. These questions aim to elicit detailed insights into the participant's skill set. Examples include queries such as, "What technical expertise do you possess? For instance, are you proficient in programming or do you enjoy editing digital images? Please provide specific details," and "Which administrative tasks do you excel at? For example, are you skilled in Word or Excel, or do you have experience in office administration? Please elaborate." Therefore, the user receives new occupation recommendations and can once again choose from the two options presented. After selecting an occupation and choosing an appropriate educational path from the three suggested pathways, a summary of the user's activity is displayed on the final page. At this point, the user has the option to save this summary as a PDF document by confirming with the PDF-Creation button.

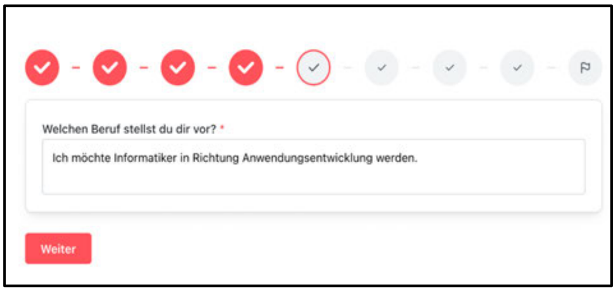

Figure 3. This figure represents the form-based variant, where the user has to respond to questions via buttons or open question text fields. With the limitation to the "Next" (German: “Weiter”) button only, progress is restricted to a forward direction and is visualized above in a progress bar where a red circle represents the current state.

2.4 Intent-based Assistant

The intent-based assistant as shown in Figure 4, is a chatbot created with the Rasa framework [17]. User input, provided as free text, is classified into a predetermined set of intents, which then dictates the execution of an appropriate dialogue. These dialogues are pre-established, with the chatbot's responses selected from a fixed pool of answers. In intent-based chatbots, interactions adhere strictly to the predetermined dialogue models, resulting in limited conversational flow and openness. In cases where the chatbot is unable to classify a user's intent, a natural language understanding fallback mechanism is activated, prompting the user with a generic message indicating that the input was not understood and requesting a rephrasing. The conversational flow within the chatbot mirrors that of the form-based assistant.

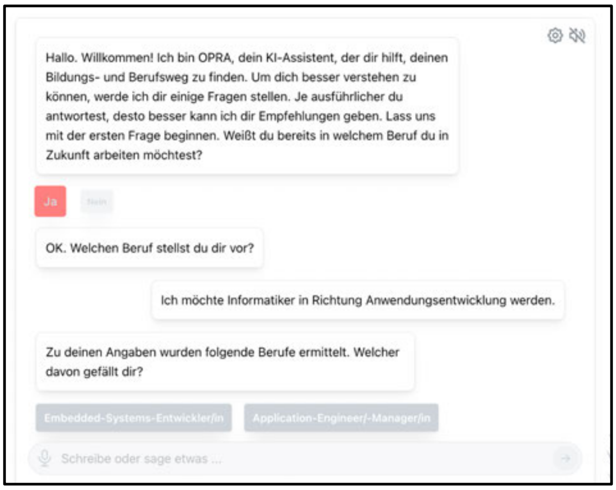

Figure 4. This figure depicts the intent-based variant, where the assistant actively initiates the conversation and supports the dialogue control flow using buttons. As illustrated, a controlled approach facilitates more precise and shorter questions.

2.5 LLM-based Assistant

As shown in Figure 5, the LLM-based assistant operates as a chatbot using the Llama-2-13B-chat-GPTQ model [18], a quantized variant of Meta's Llama2 model [19] specifically designed for chat assistant applications. Unlike deterministic models, the LLM-based assistant employs a generative approach, where understanding user intent and generating responses are not predetermined but influenced through prompt engineering. Consequently, conversations with the assistant are highly open-ended, allowing for a natural flow of dialogue between users and the chatbot. However, this lack of conversational guidance poses a significant challenge, potentially leading conversations in a direction that is not useful for recommending educational paths. To mitigate this situation, guiding prompts are strategically integrated into the user instruction, directing the dialogue towards specific topics such as preferred locations for work or study at a predefined step in the conversation. These prompts facilitate the inclusion of occupation and educational path recommendations in relevant sections of a conversation. Furthermore, besides these recommendation options, there is also the alternative, where prompts can direct an open dialogue towards a conclusion at a particular point in the conversation using predetermined steps.

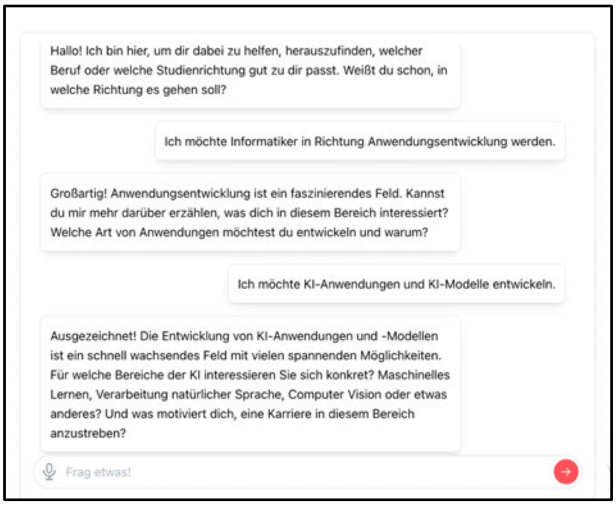

Figure 5. This figure displays the LLM-based variation, where a chat window is displayed on the right side. The conversation is actively started by the chatbot with a question of whether the user already knows a study direction or occupation. Button support is limited to recommendation results and as depicted, the questions in this variant are more open and longer.

2.6 Behavioral Measurement

To assess the performance of the assistants, a range of metrics were gathered as described next. This included calculating the average length of user input and the average response length of the assistant. Additionally, each interaction with the assistant was timestamped, enabling the calculation of the total conversation duration and the time spent on each text input. Furthermore, the recommendations made by the assistant and those selected by the user were logged, encompassing both occupations and educational pathways. The tracking also extended to identifying instances where alternative occupations were requested and recommended.

2.7 Self-Report Measurement

During the study, participants filled out two questionnaires. The initial questionnaire, completed before interacting with the assistant, aimed at responses regarding previous experiences with similar systems and attitudes towards technology and technology acceptance [20]. The second questionnaire, which was answered after their interaction with the assistant, explored participants' experiences and perceptions of the assistant using the bot usability scale (BUS-15) [21], participants' trust in the assistant using a questionnaire adapted from [22], variables perceived competence and perceived autonomy of the self-determination theory (SDT) using the technology-based experience of need satisfaction-interface (TENS-Interface) scale [23], questions regarding anthropomorphization adapted from [24] and explainability of the assistant using the system causability scale [25].

3 RESULTS

The pre-test study provided valuable insights into how pupils interact with the three different assistants tested. A total of eight pupils and one teacher participated in the study. Out of the eight pupils, six pupils completed both the questionnaires and the testing phase. Completion of testing was defined by a pupil receiving an educational path recommendation at the end of their interaction with the assistant. Three participants tested the LLM-based variant, while only one completed everything. For the intent-based variant, two participants tested it and one completed it, while four tested and completed everything for the form-based variant. The incomplete testing in some cases was partially due to technical issues that necessitated reloading the assistant. The IDs were anonymized, so that it is not known which assistant the teacher tested.

During the user testing, several observations were made regarding the interactions of the pupils with the assistants. These observations, along with reflections on pupil's feedback and insights from the data collected, are discussed below.

3.1 Lessons Learned from Observations

During the testing phase, the researchers observed the pupils and recorded their behavior.

Explorative behavior: Throughout the testing phase, the pupils highly interacted with each other, displaying curiosity about each other's experiences. They frequently looked at each other's screens and exchanged insights and experiences during the testing session. Additionally, the LLM-based version also triggered curiosity among the pupils; some attempted to "challenge" the assistant by providing prompts to elicit unexpected responses. Despite ChatGPT being a recent tool that has similarities to the LLM-based version, the pupils seemed to be familiar with this conversational format. They explored the LLM-based assistant in diverse ways; for example, one pupil asked about the assistant's guess regarding their favorite musician. These behaviors could potentially influence individual responses and the overall test outcomes. To mitigate this in future sessions, improving the initial instructions will be crucial, emphasizing the importance of independent interaction with the assistants, specifically focused solely on career-related participants.

Reading and Comprehension Abilities: Moreover, a considerable disparity in reading and comprehension abilities was noted among the pupils. While some required assistance with specific terms such as 'causality,' others found the volume of reading required for completing the questionnaires and engaging with the chatbot challenging. This observation underscores the need for revising the language used in both the questionnaires and the chatbots to ensure the study is more inclusive and accessible to all participants, regardless of their literacy levels. Additionally, having teachers or other assistants present during the user test to provide additional support is beneficial.

Study Scalability: During the pre-test phase, pupils used desktop computers. However, for a larger-scale study, the use of tablets seems to be more appropriate, as many schools now incorporate them into their curricula, familiarizing pupils with their operation. With tablets, pupils can also easily access web pages via QR codes, potentially accelerating the navigation through various links.

User ID Reuse: Another lesson learned from the pre-test involved the utilization of user IDs. It was observed that certain pupils reused the same IDs to reload web pages and to restart the experiment. To maintain data integrity, the ability to reload multiple times should be prevented and consequently limited to single use.

Accessible Interface: Lastly, incorporating speech-to-text and text-to-speech capabilities could improve accessibility. These features would be particularly advantageous for pupils facing challenges with reading and writing, guaranteeing that all participants can interact with the assistants effectively and autonomously.

3.2 Lessons Learned from Feedback

The pupils' feedback brought up various insights and offered valuable suggestions for further adjustments to the assistants.

Recommendation Modification: One participant suggested implementing a "regenerate" button to enable users to receive alternative recommendations if the initial ones were unsatisfactory. Additionally, several pupils remarked that they found the provision of only two recommendations at the end insufficient.

Survey Critique: Another valuable observation was that the post-survey was not appropriately tailored to the form-based assistant. Certain questions in the post-survey are related to the quality of the conversation. Given that the form-based assistant does not involve conversation but rather requires only the completion of answer boxes, this discrepancy proved frustrating for the pupils.

Intent-Based Responsiveness: Feedback regarding the intent-based chatbot highlighted, that pupils felt a lack of response to their inputs, particularly when posing questions during the conversation. The intent-based assistant operates with predetermined questions and cannot dynamically respond to inputs that do not trigger a specific intent.

LLM-based Assistant Complexity: A comment concerning the LLM-based chatbot noted that its questions were challenging to answer, particularly when individuals were uncertain about their desired career or study direction.

Recommendation Quality: Among the pupils, there were varying perceptions of the quality of the provided recommendations. While some found the recommendations to be well-suited to their interests, others expressed dissatisfaction because the recommendations did not align precisely with their personal preferences and desires.

Teacher Feedback: The teachers involved displayed significant interest in the workshop concept, highlighting that the information provided about LLMs was particularly insightful for them. With the growing integration of AI tools like ChatGPT into educational settings, teachers are faced with inquiries about how to thoughtfully use the new technology. The feedback underscored a pressing need for educational initiatives to equip teachers with knowledge about the challenges and possibilities of LLMs, enabling them to effectively integrate this into their curriculum.

3.3 Lessons Learned from Data

Different Time: The testing of the three assistants differed with respect to the average time needed for completion. While the pupils interacting with the form-based assistant were finished quickly, the interaction with the intent-based and the LLM-based assistant took more time. To keep the pupils busy, they were able to test one of the other assistants out of interest if they were finished early. For optimal testing, it is recommended to align the length of the interaction time.

Quality of recommendation: The recommendations often failed to fit to the interests and wishes the participants mentioned during the interaction of the assistants. For instance, one pupil explicitly stated that he/she wants to record videos as an occupation, but the recommendations were 'Project Manager for Workshops for People with Disabilities' and 'Nutritionist', both not related to that. At other times the recommendations were more closely related, for instance, one pupil was talking about his interest in mathematics and the recommendation was 'Calculator' and 'Machine Learning Engineer'. However, the quality of recommendation is something that we need to improve further. This depends largely on a good and diverse database of occupations and educational paths, as well as on powerful matching algorithms.

Text length differences: The average response length of the assistants varied among each other. While for the LLM-based assistant, the average response length was 370 characters, for the intent-based assistant it was around 170 characters, and for the form-based assistant the average length was fixed to 37 characters, as there was no variation. Differences in text length may influence the results, so it is recommended to consider reducing differences of text length in order to exclude any confounding variables.

4 CONCLUSIONS

In summary, it can be asserted that the concept of the workshop, coupled with a pre-test in schools, offers substantial benefits. These include not only valuable feedback and ongoing education but also the acquisition of user data. The former encourages pupils to participate in pre-test, thereby granting both pupils and teachers a deeper understanding of the conducted experiment. The latter facilitates testing in an uncontrolled environment (field test environment) and enhances awareness among pupils about the significance of the captured data.

The pre-test was successfully used as a framework to identify and address weaknesses and issues related to testing the assistants, including pupils' communication and interaction as well as the repetitive use of assistants. From a data perspective, this provides an opportunity to anticipate trends or tendencies in a specific direction.

The limited size of participants makes it difficult to draw any conclusions about the performance and differences of the three assistants. A larger study therefore needs to be conducted to verify trends. However, from the increased interaction time with the chatbots and the longer average responses of the LLM-based assistant, it can be assumed that the responses were adapted to the user input. The fact that LLMs are able to adapt to user preferences and generate responses in an appropriate manner [6] suggests that the conversation with the LLM-based assistant is perceived as more human-like and will have a positive impact on trust.

ACKNOWLEDGEMENTS

This research is supported by the MERLOT (MarkEtplace foR LifelOng educaTional dataspaces and smart service provisioning) project (number 68GX21008A), which is co-funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) in Germany.

REFERENCES

[1] A. Behr, M. Giese, K.H. Teguim, K. Theune. “Dropping out from Higher Education in Germany: An Empirical Evaluation of Determinants for Bachelor Students”, *Open Education Studies*, vol. 2, no. 1, pp. 126-148, 2020. Retrieved from https://doi.org/10.1515/edu-2020-0104.

[2] M. Suhlmann, K.Sassenberg, B.Nagengast, U. Trautwein. “Belonging Mediates Effects of Student-University Fit on Well-Being, Motivation, and Dropout Intention”, *Social Psychology*, vol. 49, no. 1, pp. 16-28, 2018. Retrieved from https://doi.org/10.1027/1864-9335/a000325.

[3] K., Viktória; D., Anca; G., Itamar. “Challenges and difficulties in career decision making: Their causes, and their effects on the process and the decision.” *Journal of Vocational Behavior*, vol. 116., pp. 103346, 2020.

[4] J. Moote, & L. Archer, “Failing to deliver? Exploring the current status of career education provision in England.” *Research Papers in Education,* vol. 33., no. 2, pp. 187-215, 2018

[5] S. Ekin, “Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices”. *Authorea Preprints*, 2023.

[6] T., Zylowski; M. Wölfel, “An NLP Analysis of ChatGPT’s Personality Simulation Capabilities and Implications for Human-centric Explainable AI Interfaces”, *In Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pp 168–177, Online. Association for Computational Linguistics., 2023. Retrieved from https://aclanthology.org/2023.icnlsp-1.17

[7] H. Alkaissi, & S. I. McFarlane, “Artificial hallucinations in ChatGPT: implications in scientific writing”, *Cureus*, vol. 15., no. 2., 2023.

[8] M. Wölfel, M. B. Shirzad, A. Reich, & K. Anderer, “Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice’s Cooperative Principles and Trust”, *Big Data and Cognitive Computing*, vol 8., no. 1, pp. 2, 2023. Retrieved from https://www.mdpi.com/2504-2289/8/1/2

[9] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., “GPT-4 technical report,” *arXiv preprint,* 2023. Retrieved from https://doi.org/10.48550/arXiv.2303.08774.

[10] Intelligent, Accessed 06 May, 2024. Retrieved from https://www.intelligent.com/new-survey-finds-students-are-replacing-human-tutors-with-chatgpt/

[11] Merlot Education, Accessed 08 May, 2024, Retrieved from https://merlot-education.eu/

[12] A. Gupta, P. Zhang, G. Lalwani, & M. Diab, “*Context-aware self-attentive natural language understanding for task-oriented chatbots”*, 2019.

[13] A. J. Simon, C. L. Gallen, D. A. Ziegler, J. Mishra, E. J. Marco, J. A. Anguera, & A. Gazzaley, “Quantifying attention span across the lifespan”, *Frontiers in Cognition*, vol. 2, 2023

[14] O*NET Interest Profiler, Accessed 08 May 2024. Retrieved from https://www.mynextmove.org/explore/ip

[15] Check-U, Accessed 08 May 2024. Retrieved from https://www.arbeitsagentur.de/bildung/welche-ausbildung-welches-studium-passt

[16] M. M. Nauta. “The Development, Evolution, and Status of Holland’s Theory of Vocational Personalities: Reflections and Future Directions for Counseling Psychology,” *Journal of Counseling Psychology*, vol. 57, no. 1, pp. 11–22, 2010. Retrieved from https://doi.org/10.1037/a0018213.

[17] Rasa, Accessed on 08 May, 2024. Retrieved from https://rasa.com/

[18] Llama-2-13B-chat-GPTQ model, Accessed on 08 May 2024. Retrieved from https://huggingface.co/TheBloke/Llama-2-13B-chat-GPTQ

[19] H. Touvron, L. Martin, K. Stone, et al. “Llama 2: Open Foundation and Fine-Tuned Chat Model,” *arXiv preprint*, 2023. Retrieved from https://doi.org/10.48550/arXiv.2307.09288

[20] J. Spilski, C. Giehl, T. Lachmann, J. Mayerl. “Attitudes towards Technologies and Technology Acceptance. An empirical study on innovations in construction craft,” *SocArXiv*, 2022. Retrieved from https://doi.org/10.31235/osf.io/pesb9

[21] S. Borsci, A. Malizia, M. Schmettow et al. “The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents”, *Personal and Ubiquitous Computing,* vol. 26, pp. 95–119, 2022. Retrieved from https://doi.org/10.1007/s00779-021-01582-9

[22] C.B. Nordheim, A. Følstad, C.A. Bjørkli. “An Initial Model of Trust in Chatbots for Customer Service—Findings from a Questionnaire Study”, *Interacting with Computers*, vol. 31, no. 3, pp. 317–335, 2019. Retrieved from https://doi.org/10.1093/iwc/iwz022.

[23] D. Peters, R.A. Calvo, R.M. Ryan. “Designing for Motivation, Engagement and Wellbeing in Digital Experience”, *Frontiers in Psychology*, vol. 9:797, 2018. Retrieved from https://doi.org/10.3389/fpsyg.2018.00797.

[24] S. Melián-González, D. Gutiérrez-Taño, J. Bulchand-Gidumal. “Predicting the intentions to use chatbots for travel and tourism”, *Current Issues in Tourism*, vol. 24(2), pp. 192–210, 2021. Retrieved from https://doi.org/10.1080/13683500.2019.1706457.

[25] A. Holzinger, A. Carrington, H. Müller. “Measuring the Quality of Explanations: The System Causability Scale (SCS)”, *KI - Künstliche Intelligenz*, vol.34(2), pp. 193–198, 2020. Retrieved from https://doi.org/10.1007/s13218-020-00636-z.