Enhancing Chatbot-Assisted Study Program Orientation[2]

Thilo I. Dieing [1], Marc Scheffler [1] and Lea Cohausz [1]

Abstract: As university dropout rates increase, implementing innovative solutions is crucial to reduce attrition. Aligning students' interests with their study programs enhances academic success, satisfaction, and retention. This paper presents a novel approach using open-source Large Language Models (LLM) and Retrieval-Augmented Generation (RAG) to develop a semi-open-domain knowledge chatbot. The chatbot generates informed responses and recommendations to diverse student queries by retrieving relevant data while maintaining ethical standards and avoiding biased responses. When testing five model combinations on 70 prompts partially from real study advisors, results demonstrate that the RAG approach with the Mixtral LLM and RoBERTa embedding model offers superior performance. Our method for handling critical user prompts further indicates a significantly improved response quality. These findings advance service-oriented chatbots in education, aiming to reduce student attrition through accurate and helpful program recommendations.

Keywords: chatbot, study program recommendation, LLM, RAG, CASPO

1 Introduction

With increasing university dropout rates of currently more than every fourth student in Germany [HHS22], scholars and practitioners are looking to find new ways to decrease attrition. One finding in this context is that the fit between students' interests and the content of the study program is particularly important for the development of long-term academic success [AR10; TR06]. Furthermore, scholars have demonstrated that a better fit can help increase the students' satisfaction with the program [Al08], their grades [EN16; TR06], as well as decrease the probability of attrition [BGD22]. Moreover, managing expectations and informing students via orientation systems can increase students' motivation and expectation of success [Me24]. Based on these findings, scholars have built multiple web-based systems to inform students about study programs, helping them orient themselves and make fitting recommendations. One specific method that has only received little attention is using chatbots to make study program recommendations. When they are used, chatbots are often rule-based, with human-coded questions and answers [Di18; FSW23; Ku23], which becomes infeasible for usually broader study orientation inquiries.

In this paper, we address this gap by introducing an open-source Large Language Models (LLM) and Retrieval-Augmented Generation (RAG) method to build a semi-open-domain

1 University of Mannheim, Chair of Artificial Intelligence, B 6, 26, 68159 Mannheim , Germany,
thdieing@mail.uni-mannheim.de, https://orcid.org/0009-0000-2256-0795;
marcscheffler@outlook.com, https://orcid.org/0009-0004-1249-3487;
lea.cohausz@uni-mannheim.de, https://orcid.org/0000-0002-6164-3988

2 The article appears in the workshop proceedings of DELFI 2024 and was presented at the EduRS workshop.

knowledge chatbot that can generate answers from the retrieved data and answer broader student questions. We further discuss approaches to make the chatbot usable in an ethical manner and limit its capacity to take political or otherwise sensitive standpoints that could harm the trust of users or bias the answers. On the basis of 70 prompts that partially come from real study advisors, we test five different model combinations. The results indicate that using filter layers first and applying the RAG approach with the Mixtral LLM and RoBERTa embedding model returns the best answers with good scores across all prompt types. The analysis further reveals that the proposed method for handling critical user prompts also improves the chatbot's answers significantly compared to chatbots that do not use the method.

We hope that the discussed approaches and results can be helpful to the ongoing investigation of service-oriented chatbots in the education sector and reduce attrition by informing students and giving them sensible recommendations.

2 Background

In recent years, the use and implementation of chatbots have seen a rise in almost all areas, e.g., in Management, Support, Education, Business, and Health Care [AM20; NC17]. In general, a chatbot can be regarded as an example of intelligent Human-Computer Interaction [BK18]. In the literature, scholars rely on four domains to distinguish chatbots [AM20; NC17]. From a knowledge perspective, a chatbot can either be open, meaning that one can ask the chatbot a variety of questions unbound by the topic, or closed, in which case the chatbot's knowledge will be limited to one predefined topic area. Furthermore, one can distinguish the chatbots by their service (interpersonal, intrapersonal, inter-agent) and goal (informative, chat-based, task-based). The fourth domain describes how chatbots function. They may be rule-based, which limits the chatbots' answers to a predefined human-coded set of questions and answers, or retrieval-based, or generate answers based on machine learning models. While these domains help distinguish chatbots, one must also keep in mind that the borders between different styles flow seamlessly.

2.1 Chatbots for Study Program Orientation

Similar to other areas, scholars have implemented a variety of chatbots to assist in the education sector. As such, scholars have, e.g., built chatbots to teach coding languages [DTZ20; Wi20], engineering [Me20], or psychology [Ha13]. A majority of these chatbots (88.88%) can be classified as rule-based chatbots being based on predefined answer and question possibilities [Ku23]. In the education-specific strand of literature, scholars have also categorized chatbots further as either teaching- or service-oriented [OA21; PDP20]. While teaching-oriented chatbots support educators as additional tutors or learning tools, service-oriented chatbots are more used to assist students in their day-to-day lives, making curriculum and study program decisions. Following this classification, we will focus on service-oriented chatbots to help in the study program orientation in the context of this study. Examples of chatbots designed for this specific task include LiSA, a chatbot created to help students in their university life [Di18] or the German bot DIAS, which offers students an information platform for their study pursuits [FSW23].

While these chatbots indicate promising results when it comes to study program orientation, they are, much like the majority of educational chatbots, mainly rule-based and closed-domain. This, however, limits the chatbot's ability to help students since the orientation process includes a wide set of possible user questions that cannot feasibly be accounted for by human-coded rules. Additionally, the rule-based method only works for smaller projects; as soon as a chatbot is expected to be used by a larger crowd, the number of human-coded rules increases exponentially, making the method unattainable. This work aims to fill this gap using open-source Large Language Models (LLM) to build a chatbot that can generate answers to make cross-domain recommendations and also answer broader study questions and fears. Such a chatbot can be categorized as a semi-open-domain, interpersonal, informative/chat-based, and generative bot. With such a configuration, the chatbot would offer a novel solution to the Chatbot-Assisted Study Program Orientation (CASPO).

3 Methodology

3.1 Building an Open-Domain Chatbot

For building the chatbot, we decided against a time- and resource-consuming fine-tuning of LLM models and instead chose to connect the LLM models to a dataset, thereby utilizing the Retrieval-Augmented Generation approach [Ga24]. Using this method, the user enters a prompt, which an embedding model then embeds. Following this, the model further compares the similarity between the user prompt and the database entries. The fixed number of entries that match the prompt best is then given to the LLM as additional information for the answer generation. Hence, the user is given factual information retrieved from the embedded data but can also ask broader questions regarding their study orientation.

In the context of this paper, we will focus on building a CASPO for study programs in Baden-Wuerttemberg, Germany[3]. Nevertheless, the proposed methodology can easily be adapted to incorporate more study programs. To construct the database, we scraped all necessary study program information from the website of the Federal Employment Agency[4]. The collected data included program names, descriptions, administrative details, and university information. This data was then embedded and saved as a database. Using the explained RAG methodology, we compared various embedding models and LLM model combinations to determine the best set for this specific chatbot task.

3 The chatbot is being built as part of a project initiated by the Ministry of Science, Research and Art of the State of Baden-Württemberg, the goal is to add the chatbot to a new study orientation system which can then be used.

4 `https://web.arbeitsagentur.de/berufenet/studienfelder`, last visited, July 2nd, 2024

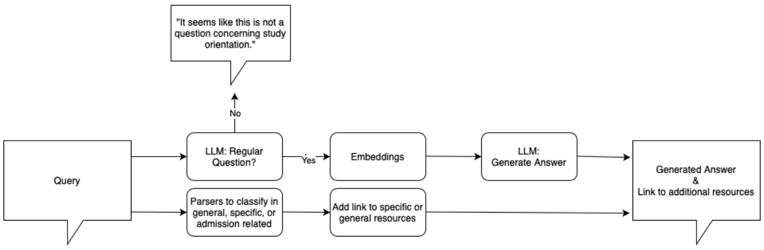

Fig. 1: Methodology to ensure the chatbot's usability and political correctness.

For the embedding models, we included a Cross English & German RoBERTa Sentence embedding model [Ma20] and a German jina embedding model [Mo24]. Both models have the advantage of performing well for the German case and for English titles of study programs held in English. In the first phase of testing, we included three different LLMs to generate answers based on embedded matches, Meta's Llama3 model [AI24], Mistralai's Mixtral 8x7B [Ji24], and the Sauerkraut LLM [VA24]. However, the first test results, analyzing the models' reasoning, demonstrated that an increased number of matches given to the LLM improved the answers, especially for broader questions, while also increasing the number of tokens the model would have to handle. Since both Llama3 and Sauerkraut support small token sizes only, we decided to exclusively further test Mixtral. Moreover, the number of matches was included as a hyperparameter, testing the LLM with 30 and 60 matches, resulting in four different chatbot configurations to evaluate.

3.2 Methods for Increasing Answer Validity and Reliability

Even though the proposed chatbot architecture introduces many novel features to CASPO, the method also has its disadvantages, which are mainly caused by the open domain. As such, scholars have demonstrated that open-domain models show clear political tendencies leaning toward a left ideology [HSW23]. Moreover, models also answer politically sensitive topics and offend users [Ba21], posing as a biased information source [Vä20]. Such biased answers could, therefore, also hurt the chatbot's effectiveness for the program orientation and may even lead to more attrition due to wrong expectations and motivations [Me24] as well as reduce the trust users have in the chatbot. For this reason, we decided to suppress some answers of the chatbot via manually included rules, resulting in a semi-open chatbot. The derived methodology for implementing this kind of chatbot is visualized in Fig. 1. The main idea is to first screen the user prompts using the LLM model instructed to detect whether the inquiry is of a sensitive or problematic nature. If so, the model will give a predefined answer. Only if the prompt is cleared to be non-problematic it is given to the embedding model and to the LLM to generate an answer. Nevertheless, utilizing the LLM for detecting critical prompts might introduce new biases from the model, as such we recommend training a custom filtering model once sufficient user queries are collected. Additionally, since LLMs are prone to hallucinate when using links, we have added an additional layer in which custom parsers scan the user prompts to add useful links from another dataset to the LLM's response. By deploying this methodology, we hope to circumvent the apparent disadvantages of open-domain chatbots and only leave the clear advantages of deploying such a model for CASPO.

3.3 Evaluation

In this study, we evaluate the four embedding/LLM combinations (jina/Mixtral30, RoBERTa/Mixtral30, jina/Mixtral60, RoBERTa/Mixtral60) and choose the best model, which we subsequently input into the discussed semi-open pipeline. A comparison of this model with the other four then shows how successful the proposed methodology is and, therefore, the usability of the chatbot. The evaluation is based on a set of 70 diverse prompts, ranging from regular prompts such as "Can I study [program] at the University of [location]?" or "I am interested in [topics], what study program can you recommend?", to explorative prompts including Reddit questions, unusual prompts, and critical prompts to see how the models react to political and offensive inquiries. In addition to prompts for these four categories, we also include a set of prompts collected by real study advisors to test how well the chatbots answer these real-world questions. The evaluation metric is coded with values of -1, meaning the answer is wrong; 0, the answer is not perfect but would suffice; and 1, the answer is correct. Based on this metric, the chatbot answers were graded by three independent coders.

4 Results

The analysis of the codings revealed an inter-coder agreement of 0.75 using Krippendorff's Alpha [HK07; Kr18], indicating a strong agreement between the coders. The individual scores were then averaged over the three coders to generate the final scores. Fig. 2 reports the average performance of the four combinations and the semi-open chatbot grouped by the discussed prompt types. The results further support the findings from the first analysis, which states that a higher number of entries given to the LLM is beneficial for answer generation. Furthermore, one can see that the results using the RoBERTa embedding model are better than those utilizing jina, with the exception of critical and regular prompts. Among the four models, the best-performing chatbot used the RoBERTa embedding model, returning 60 matched entries to the Mixtral LLM. This model performed best for explorative questions followed by regular and study advisor prompts, with average scores above 0.75. Nevertheless, the model demonstrates great difficulties and uncertainty for critical prompts.

Based on the comparison of the four chatbot combinations, we implemented the derived methodology for a semi-open chatbot with the RoBERTa/Mixtral60 architecture.

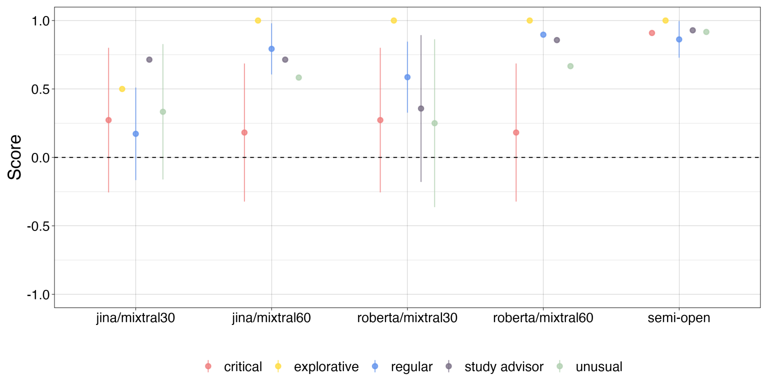

Fig. 2: Evaluation of different chatbot combinations. Note: Results report the average performance for each prompt type on a scale from -1 to 1 and the 95% confidence intervals as a measure of uncertainty.

While there are no notable changes in answers on explorative, regular, or study advisor prompts, one can see that the semi-open chatbot performs significantly better for unusual and critical prompts with an average score of 0.909, indicating the validity and reliability of the applied filtering. Hence, the results demonstrate that using the semi-open chatbot schema leads to almost perfect answers in all five categories. Since the final goal is to deploy the chatbot to be used by people interested in studying, we have uploaded the whole bot on Huggingface (https://huggingface.co/spaces/nestole/Chaetti), where we will continue with further evaluations and tests.

5 Conclusion

This work presented a novel method to create a semi-open chatbot for study program orientation and recommendation. A chatbot of such an architecture is a more scalable solution compared to current chatbots in this strand of literature. A detailed evaluation of different models and configurations has shown that filtering user prompts at the beginning and then using the RAG approach with the Mixtral LLM and RoBERTa embedding model returned the best answers with scores above 0.75 for all prompt types. In the next step, we recommend testing the developed chatbot in the field via a user survey in which we will also explore its acceptance by different demographics [BIS24; PDP20]. We further recommend investigating the effects of such chatbots on users as well as comparing different chatbot modalities such as voice- and text-based effects [RBH22]. Lastly, while the derived approach has demonstrated its merit, we believe it should be further tested to improve the approach with additional layers by including models that can, e.g., detect the user intent [ASA23] to make even better-tailored study program recommendations.

References

[AI24] AI@Meta: Llama 3 Model Card. 2024, URL: `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

[Al08] Allen, J.; Robbins, S. B.; Casillas, A.; Oh, I.-S.: Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. Research in Higher Education 49, S. 647–664, 2008.

[AM20] Adamopoulou, E.; Moussiades, L.: An overview of chatbot technology. In: IFIP international conference on artificial intelligence applications and innovations. Springer, S. 373–383, 2020.

[AR10] Allen, J.; Robbins, S.: Effects of interest–major congruence, motivation, and academic performance on timely degree attainment. Journal of counseling psychology 57 (1), S. 23, 2010.

[ASA23] Assayed, S.; Shaalan, K.; Alkhatib, M.: A chatbot intent classifier for supporting high school students. EAI Endorsed Transactions on Scalable Information Systems 1, 2023.

[Ba21] Bang, Y.; Lee, N.; Ishii, E.; Madotto, A.; Fung, P.: Assessing Political Prudence of Open-domain Chatbots, 2021, arXiv: `2106.06157 [cs.CL]`, URL: `https://arxiv.org/abs/2106.06157`.

[BGD22] Bäulke, L.; Grunschel, C.; Dresel, M.: Student dropout at university: A phase-orientated view on quitting studies and changing majors. European Journal of Psychology of Education 37 (3), S. 853–876, 2022.

[BIS24] Bilquise, G.; Ibrahim, S.; Salhieh, S. M.: Investigating student acceptance of an academic advising chatbot in higher education institutions. Education and Information Technologies 29 (5), S. 6357–6382, 2024.

[BK18] Bansal, H.; Khan, R.: A review paper on human computer interaction. International Journal of Advanced Research in Computer Science and Software Engineering 8 (4), S. 53, 2018.

[Di18] Dibitonto, M.; Leszczynska, K.; Tazzi, F.; Medaglia, C. M.: Chatbot in a Campus Environment: Design of LiSA, a Virtual Assistant to Help Students in Their University Life. In (Kurosu, M., Hrsg.): Human-Computer Interaction. Interaction Technologies. Springer International Publishing, Cham, S. 103–116, 2018, ISBN: 978-3-319-91250-9.

[DTZ20] Daud, S. H. M.; Teo, N. H. I.; Zain, N. H. M.: Ejava chatbot for learning programming language: Apost-pandemic alternative virtual tutor. International Journal 8 (7), S. 3290–3298, 2020.

[EN16] Etzel, J. M.; Nagy, G.: Students' perceptions of person–environment fit: Do fit perceptions predict academic success beyond personality traits? Journal of Career Assessment 24 (2), S. 270–288, 2016.

[FSW23] Fersch, M.-L.; Schacht, S.; Woldai, B.: Exploring AI in Education: A Quantitative Study of a Service-Oriented University Chatbot. In: The Paris Conference on Education 2023: Official Conference Proceedings. S. 2758–0962, 2023.

[Ga24] Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H.: Retrieval-Augmented Generation for Large Language Models: A Survey, 2024, arXiv: `2312.10997 [cs.CL]`, URL: `https://arxiv.org/abs/2312.10997`.

[Ha13] Hayashi, Y.: Learner-support agents for collaborative interaction: A study on affect and communication channels. 2013.

[HHS22] Heublein, U.; Hutzsch, C.; Schmelzer, R.: Die Entwicklung der Studienabbruchquoten in Deutschland. DZHW Brief (05|2022), 2022, DOI: 10.34878/2022.05.dzhw_brief, URL: https://doi.org/10.34878/2022.05.dzhw_brief.

[HK07] Hayes, A. F.; Krippendorff, K.: Answering the Call for a Standard Reliability Measure for Coding Data. Communication Methods and Measures 1 (1), S. 77–89, 2007, DOI: 10.1080/19312450709336664, eprint: https://doi.org/10.1080/19312450709336664, URL: https://doi.org/10.1080/19312450709336664.

[HSW23] Hartmann, J.; Schwenzow, J.; Witte, M.: The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv preprint arXiv:2301.01768, 2023.

[Ji24] Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; Sayed, W. E.: Mixtral of Experts, 2024, arXiv: 2401.04088 [cs.LG], URL: https://arxiv.org/abs/2401.04088.

[Kr18] Krippendorff, K.: Content analysis: An introduction to its methodology. Sage publications, 2018.

[Ku23] Kuhail, M. A.; Alturki, N.; Alramlawi, S.; Alhejori, K.: Interacting with educational chatbots: A systematic review. Education and Information Technologies 28 (1), S. 973–1018, 2023.

[Ma20] May, P.: Cross English German RoBERTa for Sentence Embeddings. 2020, URL: https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer.

[Me20] Mendez, S.; Johanson, K.; Martin Conley, V.; Gosha, K.; A Mack, N.; Haynes, C.; A Gerhardt, R.: Chatbots: A tool to supplement the future faculty mentoring of doctoral engineering students. International Journal of Doctoral Studies 15, 2020.

[Me24] Merkle, B.; Bürkle, H.; Janke, S.; Karst, K.: Change my mind. Zeitschrift für Pädagogische Psychologie, 2024, DOI: 10.1024/1010-0652/a000379, eprint: https://doi.org/10.1024/1010-0652/a000379, URL: https://doi.org/10.1024/1010-0652/a000379.

[Mo24] Mohr, I.; Krimmel, M.; Sturua, S.; Akram, M. K.; Koukounas, A.; Günther, M.; Mastrapas, G.; Ravishankar, V.; Martínez, J. F.; Wang, F. et al.: Multi-Task Contrastive Learning for 8192-Token Bilingual Text Embeddings. arXiv preprint arXiv:2402.17016, 2024.

[NC17] Nimavat, K.; Champaneria, T.: Chatbots: An overview types, architecture, tools and future possibilities. Int. J. Sci. Res. Dev 5 (7), S. 1019–1024, 2017.

[OA21] Okonkwo, C. W.; Ade-Ibijola, A.: Chatbots applications in education: A systematic review. Computers and Education: Artificial Intelligence 2, S. 100033, 2021.

[PDP20] Pérez, J. Q.; Daradoumis, T.; Puig, J. M. M.: Rediscovering the use of chatbots in education: A systematic literature review. Computer Applications in Engineering Education 28 (6), S. 1549–1565, 2020, DOI: https://doi.org/10.1002/cae.22326, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cae.22326, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cae.22326.

[RBH22] Rzepka, C.; Berger, B.; Hess, T.: Voice assistant vs. Chatbot–examining the fit between conversational agents' interaction modalities and information search tasks. Information Systems Frontiers 24 (3), S. 839–856, 2022.

[TR06] Tracey, T. J.; Robbins, S. B.: The interest–major congruence and college success relation: A longitudinal study. Journal of vocational behavior 69 (1), S. 64–89, 2006.

[Vä20] Väänänen, K.; Hiltunen, A.; Varsaluoma, J.; Pietilä, I.: CivicBots–Chatbots for supporting youth in societal participation. In: Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3. Springer, S. 143–157, 2020.

[VA24] VAGOsolutions: SauerkrautLM-Mixtral-8x7B-Instruct. 2024, URL: `https://huggingface.co/VAGOsolutions/SauerkrautLM-Mixtral-8x7B-Instruct`.

[Wi20] Winkler, R.; Hobert, S.; Salovaara, A.; Söllner, M.; Leimeister, J. M.: Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In: Proceedings of the 2020 CHI conference on human factors in computing systems. S. 1–14, 2020.