ADAPTING A STUDENT-FACING CHATBOT TO THE NEEDS OF FIRST-GENERATION STUDENTS: A USER EXPERIENCE STUDY

Maria K Wolters[2,3], *Tatjana Kukic*[1], *Stefan Hillmann*[1]
[1]*Technische Universität Berlin, DE* [2]*University of Edinburgh, UK* [3] *OFFIS e.V., DE*
stefan.hillmann@tu-berlin.de

Abstract: First-generation students often face challenges in navigating university structures due to a lack of familial academic experience. The presented study investigates the adaptation of the student-support chatbot CHATU to better serve FGS at Technische Universität Berlin. In a two-stage evaluation, we first assessed the usability and effectiveness of the existing CHATU chatbot with FGS participants, revealing below-average usability ratings in perspicuity and dependability, as well as negative ratings in attractiveness, efficiency, stimulation, and novelty. Based on these findings, we developed CHATU-RAG, integrating Generative AI for improved interaction and adaptability. While perceived usability improved significantly, actual task success declined. User impression of the chatbot increased substantially ($p<0.0005$), highlighting the trade-offs between AI-driven interaction improvements and reliable information retrieval in an academic support system.

1 Introduction and Related Work

First-generation students (German term: Erstakademiker*innen) are less likely to be familiar with university life and success strategies than those students with a university graduate family member. In this paper, we investigate in a two-stage study (N=35) how CHATU, a chatbot that supports students in navigating their studies, might be redesigned to best support first-generation students.

The rest of this section defines first-generation students (FGS), elaborates on their prevalence in the academic system as well as their pain points, and provides an overview of CHATU. Section 2 and 3 describe the methodology of two studies on FGS using two different chatbot versions. Finally, Section 4 discusses the results of the presented studies.

1.1 First-Generation Students

First-generation students [1, p. 18] are students who are the first in their immediate family to pursue higher education. This means that neither of their parents, or primary caregivers, has obtained a university degree. In contrast, continuing-generation students (CGS) come from families where at least one parent has completed higher education. The distinction between FGS and CGS is important because it highlights structural and social disparities in educational attainment and access to academic resources. The definition of FGS varies slightly across different studies and educational systems . In some contexts, the term applies strictly to students whose parents lack any form of tertiary education, while in others, it includes students whose parents may have attended but not completed a university degree. The classification of FGS is significant in educational research as it helps in understanding the unique challenges and barriers these students face.

The proportion of FGS varies across countries and institutions, often depending on broader socio-economic factors and national educational policies. In Germany, approximately 47% of

university students starting their studies considered FGS (coming from non-academic family backgrounds) [2, extracted from Fig. 20, p. 88]. On the other hand, the probability of studying at a university is only 27% for pupils with a non-academic background, compared to 79% of those with an academic background [2, p. 13]. Finally, only 40.74% of FGS get a master's degree, while 54.43% of CGS achieve this (own calculation based on [2, p. 13]).

In other European countries [3, p. 115] and the United States [4], the percentage of FGS ranges from 20% to over 40%, depending on the institutional definitions and survey methodologies.

Many FGS lack the professional connections that their continuing-generation peers might have through family or social networks. In the first semesters, they often struggle with navigating the university system, which can be overwhelming due to unfamiliar administrative processes, academic expectations, and cultural differences. Without prior exposure to higher education environments, they may lack essential knowledge about study techniques, time management, and how to access academic resources. As FGS have no such experience in their family background, it is more difficult to get advice about the named factors and in need of other sources of information and knowledge. However, when using such, the problem of a limited study system-specific vocabulary can hinder efficient information retrieval.

1.2 CHATU Chatbot

CHATU[1] is a chatbot assistant designed to support university students in navigating their academic journey. Developed at the Technische Universität Berlin (TU), CHATU integrates multiple modules to provide structured guidance on university-related topics. These modules allow users to ask questions about general topics related to studying at TU Berlin (MENTOR), information about courses and modules (MOSES), and questions which have been directed to TU's Office of Student Affairs (SEKRETARIAT). Details on the orchestration of the modules are provided in [5], while SEKRETARIAT is introduced in [6] and its general usability evaluation is described in [7].

The module SEKRETARIAT answers questions that are usually handled by the Office of Student Affairs team. These are the questions about application and enrollment at TU, academic leave of absence, semester fees, language certificates, and similar. MENTOR provides general information about TU Berlin, like the meaning of special terms (Asta, QISPOS), where to find a cafeteria, or what an examination board does. The conception and evaluation of these CHATU modules were led by the goal to support all students at TU in the same manner, without focusing on features which might be of benefit for FGS.

In the following, we present two studies aiming to identify chatbot features which help to improve the use of chatbots like CHATU. In the first study we evaluated CHATU with FGS. On the basis of participants' feedback, we revised CHATU to add Generative AI functionality (CHATU-RAG). This allows a more conversation-like interaction and can adapt to the user's vocabulary. CHATU-RAG was evaluated in the second study.

2 Study 1: Evaluation of CHATU

2.1 Method

In order to determine the redesign requirements, we first evaluated the original chatbot, CHATU, with FGS. Based on the findings of this evaluation, we created a second chatbot, CHAT-RAG (c.f. Section 3) and evaluated it with FGS using the same procedure as for CHATU. Most partic-

[1] `https://chatu.qu.tu-berlin.de`

ipants were recruited through the TU Berlin experiment participation pool, others through personal contacts. All participants were compensated for their time with EUR 15, which is above the German hourly minimum wage. The study did not require a statement of the Ethics Committee of the Faculty IV, TU Berlin (decision reference number 287, based on a self-disclosure questionnaire).

2.1.1 Procedure

After consent, participants completed a demographics questionnaire with information about age, gender, student status, education, work commitments, and care commitments. Student status questions included whether the participant was still a student, whether they studied full time or part time, what degree they were studying, whether they were a student at TU Berlin. Technology affinity was measured using the nine item Affinity for Technology Interaction scale (ATI, [8]). In a structured interview, participants were asked about their family background, their knowledge of academic terminology, sources of support for applying to university, use of chatbots, and attitude to chatbots.

Next, participants used the chatbot in four information seeking scenarios and rated perceived task success after each interaction on a binary scale (yes / no). The scenarios were selected from a set of seven possible scenarios, which was distributed across participants using a pattern that ensured a roughly equal number of responses.

Having completed all four information seeking tasks, participants rated the chatbot using the User Experience Questionnaire (UEQ, [9]) and a questionnaire adapted from the ITU-T Rec. P.852 [10]. The UEQ is a standard usability assessment that yields six ratings, attractiveness (overall impression), perspicuity (ease of familiarisation), efficiency (time and effort), dependability (predictability), stimulation (excitement, fun) and novelty (innovative, eye catching). The adapted ITU-T questionnaire omitted judgements of overall quality and dialogue flow. The dimensions system-provided information (PI), communication with the system (CS), system behaviour (SB), user impression of the system (UIS) and acceptance (AC) were assessed using selected items from the full questionnaire. Four items assessing overall usability (US) were added. Finally, participants were debriefed about their experience using a semi-structured interview.

2.1.2 Scenarios

All scenarios, summarised in Table 1, were based on common tasks for students. While scenarios 1, 2, 4, and 6 were similar to the tasks used in [6], scenarios 3, 5, and 7 were new. All scenarios were described using a picture and a textual description of the information needs. Task success was established by reviewing the log files and scored as full (all relevant information retrieved), partial (some relevant information retrieved), or none (no relevant information retrieved).

2.1.3 Analysis

UEQ data were analysed using the Excel spreadsheets provided at `https://www.ueq-online.org`. All other statistical analysis was conducted using R 4.2.3 (Shortstop Beagle) and the packages tidyverse (data processing), psych (scale evaluation), and coin (non-parametric statistics). Significance of differences between studies were assessed using Fisher's exact test for demographic data, the paired version of the Wilcoxon-Mann-Whitney test for differences in performance on scenarios, and the Kruskal-Wallis test for differences in usability ratings. Interview findings were summarised narratively based on researcher memos and transcripts.

Table 1 – Scenarios and Success Rates. For each scenario, we provide a description, the number of trials, perceived success rate, and the objectively assessed rate of full and partial success. N_Q: Scenarios reported in questionnaire. N_L: Scenarios found in log files. Percentages are rounded to the nearest integer

Scenario	CHATU					CHATU-RAG				
	N_Q	Perceived	N_L	Full	Partial	N_Q	Perceived	N_L	Full	Partial
1: Replacement student card—application and cost										
	10	100%	11	100%	0%	9	100%	9	33%	11%
2: Semester fees—amount and payment information										
	11	100%	11	100%	0%	11	73%	11	64%	36%
3: Change of address										
	10	100%	10	100%	0%	10	100%	11	45%	27%
4: Application for Masters—documents and deadlines										
	12	83%	12	58%	33%	11	91%	11	27%	45%
5: *Semesterticket*—existence, cost, obtaining one, whether compulsory										
	12	66%	11	9%	73%	11	73%	10	50%	20%
6: Academic calendar—key deadlines and teaching vacations										
	12	83%	11	100%	0%	11	64%	10	20%	20%
7: Programme handbook—what it is and where to find it										
	10	90%	11	80%	20%	9	78%	7	29%	29%

2.2 Results

19 people participated in Study 1. On average, a session took around 50 minutes. Selected participant demographics are presented in Table 2. The mean ATI score was 3.9 (SD=1.1, Cronbach's α=0.92). First generation students mostly sought help with navigating academic life online or from friends. Chatbots were a widely used tool, but human contact was perceived to be vital for resolving open questions.

2.2.1 Usability

For reasons of space, we focus on task success (actual and perceived) and the UEQ scores. Participants were uniformly successful in scenarios 1–3. Scenarios 4, 6, and 7 were more difficult, with Scenario 5 being the most challenging. Direct comparisons between success rates are unfortunately difficult due to a few mismatches between user ratings and logs. For scenarios 5 and 6, one log each was lost, while for one perceived task success rating is missing for scenarios 1 and 7. Nevertheless, we see an interesting trend. Except for one scenario (6, academic calendar), perceived task success was equal to or higher than actual task success. This is particularly evident for Scenario 5 (*Semesterticket*[2]), where two thirds of participants considered themselves successful, but logs showed only one truly successful attempt.

In terms of UEQ scores, CHATU was rated below average in perspicuity and dependability, and badly in the other four dimensions (c.f. Figure 1a). All subscales were reliable, with Cronbach's α ranging from 0.74 (dependability) to 0.89 (perspicuity). CHATU's assessment on the ITU scales was also neutral to slightly negative (c.f. Table 3). All scales were reliable, with Cronbach's α ranging from 0.77 to 0.86, except for CS with α=0.56.

[2]A *Semesterticket* is flat fee for local public transport for the entire semester, which is commonly offered to students at German universities.

Table 2 – Selected demographics for Study 1 and Study 2: Gender, age group, student status, whether the participant is currently a student at the Technical University of Berlin, highest level of education achieved. For assessing differences in demographics between studies, Fisher's exact test was used. All percentages are rounded to the nearest integer.

		Study 1 (N=19)		Study 2 (N=17)		Sig.
		N	%	N	%	
Gender	female	11	58%	8	47%	p>0.5
	male	7	37%	9	52%	
	no answer	1	5%	0	0%	
Age Group	18–25	2	11%	2	12%	p>0.9
	25–34	8	42%	9	53%	
	35+	9	47%	6	36%	
Student Status	yes, at TU Berlin	8	42%	8	47%	p>0.4
	yes, not at TU Berlin	7	37%	3	18%	
	former student	4	21%	6	35%	

Table 3 – Mean, standard deviation, and Cronbach's α for the adapted ITU scales for CHATU and CHATU-RAG. Significance of difference between studies was assessed using the Kruskal-Wallis test. PI: provided information, CS: communication with the system, SB: system behaviour, UIS: user's impression of the system, AC: acceptance, US: additional usability items not from ITU.

Subscale	Study 1			Study 2			Sig.
	M	SD	α	M	SD	α	
PI	0.0	0.7	0.86	0.3	0.5	0.54	p<0.05
CS	0.7	0.8	0.56	1.2	0.4	-0.16	p>0.08
SB	-0.3	0.7	0.81	0.1	0.6	0.60	p<0.005
UIS	-0.5	0.8	0.86	0.6	0.7	0.82	p<0.0005
AC	-0.3	0.8	0.77	0.0	0.8	0.76	p<0.05
US	0.2	0.9	0.86	0.8	0.9	0.70	p<0.01

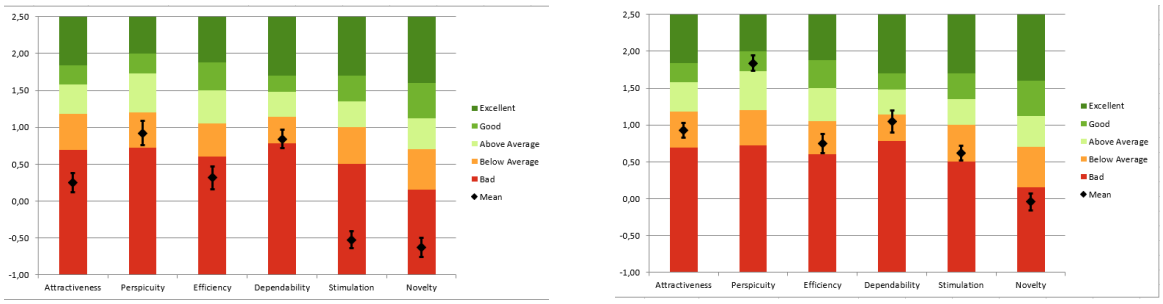
2.2.2 Narrative Summary of Interview Data

Participants appreciated CHATU's fast responses to simple questions, efficient access to information through provided links, and its minimalist design. However, some participants found the links provided too generic and were bothered by CHATU's interaction style. They felt that it was limited, generic, and impersonal. Participants also reported problems when looking for more complex information and would have preferred to contact a human in those cases. Overall, participants preferred a ChatGPT-style interface to the baseline of brief text and links provided by CHATU.

3 Study 2: Evaluation of CHATU-RAG

We created a second, retrieval augmented generation (RAG)-based version for Study 2. The RAG chatbot's engine was Mistral-7B-Instruct-v0.2 [11], the user interface was created using Open Web UI [12]. We enriched the data set with documents that included required information for all seven scenarios, data on news, and information on withdrawing from study. At the start of each interaction, CHATU-RAG provided sample prompts for common questions. Fallback messages were implemented in case of problems. The interface is shown in Figure 2.

18 people took part in the evaluation of CHATU-RAG. One participant was excluded due to invalid data. Table 2 shows demographic data for the remaining 17 participants. The average study duration was 43 minutes. Using Fisher's exact test, we established that demographics for both evaluations were similar (c.f. Table 2). The mean ATI score was 3.8 (SD=1.1, Cronbach's



(a) CHATU UEQ Scores

(b) CHATU RAG Scores

Figure 1 – UEQ scores for the six dimensions attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty (from left to right). The stacked bar charts show cut-offs for quality, ranging from excellent (top) to bad (bottom). Mean ratings are superimposed on the bar charts.

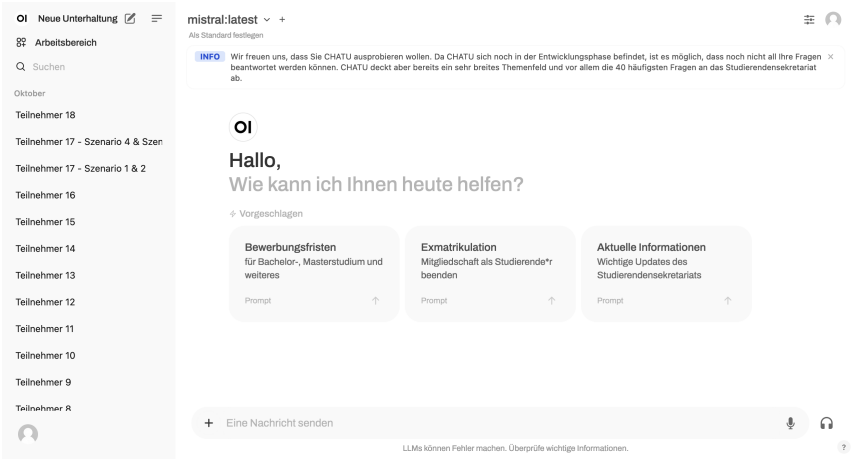


Figure 2 – Screenshot of the CHATU RAG chatbot from Study 2

α=0.93), which was similar to Study 1 participants (Kruskal-Wallis $\chi^2(1)$=0.091, p<0.76). Six participants (35%) had already participated in Study 1, eleven participants (65%) had not.

3.1 Results

3.1.1 Usability

While the difference between chatbots in terms of perceived task success was not significant (Paired Asymptotic Wilcoxon-Mann-Whitney test, Z(1)=0.851, p<0.4), full task success was worse (Z(1)=2.132, p<0.04). The only exception to that general trend was Scenario 5 (Semesterticket), where CHATU-RAG performed substantially better. For a full comparison by scenario, see Table 1.

In contrast, UEQ scores improved, as the benchmarking analysis in Figure 1b shows. Perspicuity improved substantially and was rated as good, which suggests that CHATU-RAG was easier to learn. All other dimensions now scored below average, except for novelty, which was still considered to be bad. Again, the scales proved reliable, with Cronbach's α between 0.65 (perspicuity) and 0.86 (dependability). We see a similar substantial improvement on the dimensions covered by the adapted ITU-T questionnaire (c.f. Table 3), especially in terms of system behaviour (SB) and user impression of the system (UIS). Surprisingly, the reliability of PI, CS, and SB declined substantially as measured by Cronbach's α while the other three scales (UIS, AC, and US) performed well.

3.1.2 Narrative Summary of Comments

Participants found CHATU-RAG easy to navigate, efficient, and easy to understand. However, it appeared impersonal to some, and other participants noted issues with empty links and content that did not directly address the question being asked. Those who had already taken part in Study 1 were asked to briefly compare both chatbots. Two noted no improvement, two saw clear improvement, one noted a clear decline in performance, and one person reported positive and negative changes.

4 Discussion and Conclusion

To the best of our knowledge, this paper is the first study of a chatbot specifically adapted to German FGS' needs. Interactive online tools are particularly important for FGS, who cannot easily obtain information about life as a student from their own families. In addition to friends, the FGS we interviewed relied heavily on online resources. Since interviewees also often used chatbots, a chatbot interface to official university resources should be a useful tool for FGS in particular.

In Study 1, we assessed how FGS rated the usability of the existing chatbot CHATU, which helps students navigate TU Berlin. In a prior study, [7], CHATU scored well on two complementary usability questionnaires. With FGS, however, CHATU performed much worse. Based on participant feedback, we implemented a RAG version, CHATU-RAG, and tested it in Study 2. CHATU-RAG was rated significantly more favourably than CHATU. While participants in [7] were already familiar with Generative AI applications, participants in these studies may already expect a ChatGPT style interface, given that the weekly number of users of ChatGPT doubled between April 2023 and April 2024 [13]. There was a clear disconnect between perceived and actual task success. While there was no difference between perceived task success for both versions of CHATU, actual task success declined for the RAG version. The reasons for this discrepancy need further investigation.

This preliminary study had several limitations. Since only 6 out of 17 Study 2 participants had already interacted with CHATU before CHATU-RAG, we did not achieve a clean within-subjects design. In addition, reliability of three out of the six ITU-T scales decreased; this might be due to bad item selection. Given insufficient numbers of participants, we did not control for demographic characteristics, experience, and prior knowledge of key terms in our analyses. Furthermore, we identified several issues with the CHATU-RAG implementation that led to missing or incorrect information.

In future work, we hope to test whether our findings can be generalised more widely through a comparative study of an improved CHATU-RAG with FGS and non-FGS in their first semester of studies.

5 Acknowledgments

Parts of the presented work and this paper have been funded by the Federal Ministry of Education and Research (Germany) and the Federal State of Berlin under grant no. 16DHBKI088 for the project USOS at Technische Universität Berlin.

References

[1] MIETHE, I., W. BOYSEN, S. GRABOWSKY, and R. KLUDT: *First Generation Students an deutschen Hochschulen*, vol. 167 of *Forschung aus der Hans-Böckler-Stiftung (HBS)*.

Nomos, Baden-Baden, 1 edn., 2014.

[2] STIFTERVERBAND FÜR DIE DEUTSCHE WISSENSCHAFT: *Hochschul-Bildungs-Report - 2022 Abschlussbericht*. 2022. URL https://www.hochschulbildungsreport.de/sites/hsbr/files/hochschul-bildungs-report_abschlussbericht_2022.pdf. Last access 01-31-2025.

[3] DE LEL, G., C. RACKÉ, D. CROSIER, A. HORVATH, D. KOCANOVA, T. PARVEVA, A. RAUHVARGERS, A. DESURMONT, M.-F. PAQUET, J. RIIHELÄINEN, and EDUCATION, AUDIOVISUAL AND CULTURE EXECUTIVE AGENCY (eds.): *The European higher education area in 2015: Bologna Process implementation report*. Publications Office, Luxembourg, 2015. doi:10.2797/99035.

[4] CATALDI, E. F., C. T. BENNETT, and X. CHEN: *First-Generation Students: College Access, Persistence, and Postbachelor's Outcomes*. Tech. Rep., National Center for Education Statistics, 2018. URL https://eric.ed.gov/?id=ED580935.

[5] GÖRZIG, P., J. NEHRING, S. HILLMANN, and S. MÖLLER: *A Comparison of Module Selection Strategies for Modular Dialog Systems*. In *Elektronische Sprachsignalverarbeitung 2023*, vol. 105, pp. 40–47. TUDpress, Dresden, 2023.

[6] HILLMANN, S., P. GÖRZIG, and S. MÖLLER: *Automatic Generation of Website-Based Multi-Turn Question-Answering Dialog Systems*. In *Elektronische Sprachsignalverarbeitung 2023*, vol. 105, pp. 48–55. TUDpress, Dresden, 2023.

[7] HILLMANN, S., P. KOWOL, A. AHMAD, R. TANG, and S. MÖLLER: *Usability and User Experience of a Chatbot for Student Support*. In *35. Konferenz Elektronische Sprachsignalverarbeitung*, pp. 22–29. TUDPress, 2024. doi:10.35096/OTHR/PUB-7076.

[8] FRANKE, T., C. ATTIG, and D. WESSEL: *A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale*. *International Journal of Human–Computer Interaction*, 2019. URL https://www.tandfonline.com/doi/abs/10.1080/10447318.2018.1456150. Publisher: Taylor & Francis.

[9] SCHREPP, M. and J. THOMASCHEWSKI: *Design and Validation of a Framework for the Creation of User Experience Questionnaires*. *Int. J. of Interactive Multimedia and AI*, 5(7), p. 88, 2019. doi:10.9781/ijimai.2019.06.006.

[10] ITU-T: *Subjective quality evaluation of text-based chatbots*. Recommendation P.852, International Telecommunication Union Telecommunication Standardization Sector, 2022. URL https://www.itu.int/rec/T-REC-P.852/.

[11] MISTRAL AI: *Mistral-7B-Instruct-v0.2*. 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2. Accessed: 2025-02-01.

[12] OPEN WEBUI: *Open webui documentation*. 2024. URL https://docs.openwebui.com/. Letzter Zugriff am 27.10.2024.

[13] SINGH, S.: *Number of chatgpt users*. 2025. URL https://www.demandsage.com/chatgpt-statistics/. Last access February 1, 2025.