Usability testing of BuddyGPT: A proprietary and integrated conversational agent to support the learning of university students

Jiawei Li

Faculty of Psychology, University of Twente

University Courses: The Usability Evaluation of Higher Education Chatbot (BuddyGPT)

First supervisor: Dr Simone Borsci

Second supervisor: Nese Baz

External advisor: Dr. Gayane Sedrakyan

External advisor: Cosmin G. Ghiauru

June 30th, 2025

Abstract

With the advancement of Education 5.0 and the increasing integration of digital tools in learning environments, AI-powered chatbots have gained growing importance in the field of education. Rooted in the principles of Industry 5.0, the student-centered design philosophy has become a core guideline for developing AI-driven educational tools. In line with this approach, ensuring the usability of AI-powered educational chatbots is an essential precondition to the delivery of high-quality tools to assist students in education. The University of Twente supported the development of BuddyGPT, an AI-powered chatbot designed to assist students in retrieving course-related information from the university educational platform, i.e., Canvas. The present work presents and discusses the results of the usability assessment of BuddyGPT in order to provide actionable design recommendations to support its future development. This study used a within-subjects design. A total of 53 participants were invited to use BuddyGPT to achieve three tasks associated with a course called Software Systems at the University of Twente. During the experience, four main variables were collected, with effectiveness (task success rate and query reformulation rate), efficiency (task completion time), satisfaction (BUS-11 scale), and recommendation intent (individual-level Net Promoter Score). Overall, BuddyGPT enabled participants to complete 77.7% of the tasks. Participants yielded an average satisfaction score of 3.49, indicating the participants were satisfied with BuddyGPT. A group-level Net Promoter Score of 14.3 reflects strong user loyalty and a high likelihood of recommending BuddyGPT. Moreover, we measured relationships among the main factors in the evaluation. For instance, results confirmed the correlation between individual-level Net Promoter Score and satisfaction; nevertheless, they did not find the expected correlation between the effectiveness and efficiency of people's performance in the tasks and their declared satisfaction. These findings offered developers fresh insights for boosting satisfaction with BuddyGPT and demonstrated its value in supporting students' learning. Finally, we presented practical design recommendations to guide the future development of BuddyGPT and other AI-driven educational chatbots, accelerating their adoption and enhancing their impact in the education sector.

1. Introduction

With the increasing demand for digital learning tools, AI-powered chatbots (from now on, chatbots) have emerged as a crucial innovation in the education sector. A chatbot can be defined as a conversation system that is capable of "interactions" with users by using natural conversational language in text (Baabdullah et al., 2022).

Notably, before the COVID-19 pandemic, traditional education was already being influenced by big data and AI (Sedrakyan et al., 2024). The occurrence of COVID-19 has accelerated the intelligent transformation of education, particularly strengthening the connection between Artificial Intelligence technology and educational practices (Sain et al., 2024). This unexpected disruption underscored the need for resilient, AI-driven learning solutions. Due to the outbreak of the COVID-19 pandemic, campuses were forced to close, and the educational model shifted from in-person teaching to online learning (Bhavya et al., 2021). A wide range of new technologies has been applied to online education, such as web conferencing platforms, artificial intelligence technologies. Additionally, the development of industrial paradigms has driven the evolution of educational theories. Within the latest educational theoretical framework, the student-centered design concept is considered the core principle in chatbot design, which means that improving the usability of chatbot design is necessary. However, Kuhail et al. (2022) have found that although these systems are intended to support learning, they seldom truly follow usability principles in their design, which often leads to a decline in students' interest in using them after interacting with the chatbot. Likewise, Plantak Vukovac et al. (2021) suggest that usability plays a crucial role in the effectiveness of chatbots in education.

The onset of the COVID-19 pandemic forced the traditional education system to undergo a significant shift, particularly toward digital formats, such as chatbots, to facilitate asynchronous learning (Daniel, 2020). In this evolving landscape, Yang and Stansfield (2022) stated that AI chatbots can help teachers organize online education and assist students in identifying areas for improvement in their knowledge structure. Ng et al. (2023) also proposed that AI technology provides more learning opportunities for students, enhances interaction between teachers and students, and enables personalized and timely feedback. According to big data collected by massive online educational platforms, this shift has made AI-driven chatbots play an increasingly important role in post-pandemic education by restructuring the teaching framework (Megahed et al., 2022). In traditional education, students often rely on teachers to deliver knowledge, with limited opportunities for independent exploration and critical thinking (Selwyn, 2013). Megahed et al. (2022) defined AI-empowered education as a modern educational model that leverages artificial intelligence technology to optimize the education process. With the support of AI technology, AI-empowered education, such as intelligent recommendation systems, conversational agents, and automated assessments, to encourage students to independently identify problems, explore solutions, and develop critical thinking and innovation skills (Megahed et al., 2022). Ng et al. (2023) also proposed that AI technology provides more learning opportunities for students, enhances interaction between teachers and students, and enables personalized and timely feedback. Thus, providing a chatbot with good usability design is becoming more and more important for education.

1.2. The Main Benefits and Challenges in Applying Chatbots in Education

With the increasing number of students, the educational system is facing significant challenges. Too many students and too few teachers make it difficult to provide personalized education for each student. However, chatbots in the field of education can help teachers to provide a personalized learning experience for students (Benotti et al., 2018). For example, chatbots can identify students' knowledge gaps, answer their questions, and provide explanations at the appropriate time (Sara et al., 2023). This can reduce teachers' workload and enable more students to receive timely assistance. Additionally, teachers can use AI-powered chatbots to analyze students' learning progress by recording interaction data between the chatbot and students (Sedrakyan et al., 2024). Based on the chatbot's analysis, teachers can adjust teaching plans in real-time and provide individualized feedback for each student. This demonstrates that the focus of learning is gradually shifting from teachers to students (Megahed et al., 2022).

In nursing training, traditional lecture-based teaching methods often lack interaction and contextual learning. However, within the framework of AI-empowered education, AI-powered chatbots can enhance nursing students' critical thinking and provide realistic practice scenarios for patient care. This, in turn, improves nursing students' self-efficacy, learning engagement, and academic performance (Chang et al., 2021). Furthermore, in the field of programming, the lack of different programming solutions for the same problem is a major issue in traditional programming education. However, using chatbots as a learning platform can offer students various programming solutions, helping them master multiple programming skills (Sun et al., 2024). Despite these advantages, the application of AI technology in education still encounters several challenges. As noted by Kuhail et al. (2022), AI-powered chatbots exhibit limitations in terms of usability. Future research could address this issue by investigating how individual performance influences chatbot effectiveness and satisfaction, thereby informing strategies to enhance usability.

1.3. Education 5.0

Over the past decade, the Industry 4.0 paradigm—centered on data-driven automation and process optimization—has evolved into Industry 5.0's human-centric and value-driven vision (Xu et al., 2021). This means that improving the usability of artificial intelligence so that it can collaborate more effectively with humans has become crucial. Industry 5.0's industrial concepts and technologies have not only transformed products but have also exerted a profound influence on other sectors, including education (Shahidi Hamedani et al., 2024). Anfoud and Alami Talbi (2024) suggest that continuous advancements in industrial technology have significantly transformed education, leading to the emergence of Education 5.0. Hongli & Wai Yie (2024) propose that Education 5.0 draws from the value of Industry 5.0 and emphasizes that technology should regard providing users with value as a key point. Thus, Education 5.0 can be defined as using digital technology to eliminate learning barriers, enhance studying ways, and create a learner-centered educational environment that involves teachers, students, and administrative staff (Ahmad et al., 2023). And Kuhail et al. (2022) stress usage of educational chatbots is important for the development of education 5.0. In Education 5.0, digital technologies—such as AI-driven chatbots—require high usability to collaborate more effectively with teachers on instructional tasks and help students learn more efficiently. Therefore, against the backdrop of Industry 5.0 and Education 5.0, it is increasingly important to study and improve the usability of chatbots in education.

1.4. Definition and measuring methods of Usability

Usability is defined as the extent to which a system or product can be used by users to achieve specified goals with effectiveness, efficiency, and satisfaction within a particular context of use (ISO 9241-210, 2018). According to ISO 9241-11 (2019), effectiveness refers to the degree to which users can accurately and completely accomplish intended tasks, while efficiency concerns the resources—such as time, effort, or other inputs—required to achieve these goals. Meanwhile, satisfaction is defined as users' physical, cognitive, and emotional responses that arise from whether the outcome of system use meets their expectations (ISO

9241-210, 2019). Together, these three dimensions provide a comprehensive framework for evaluating the usability of chatbots.

Nowadays, the Bot Usability Scale (BUS)-11 scale was developed by Borsci et al. (2024) in order to assess users' satisfaction with the chatbots, and it is composed of 4 factors: accessibility, functional interactive conversation, privacy, and responsiveness. The BUS-11 provides a good validity and balance between reliability and completeness of information compared to other similar scales (Borsci et al., 2022).

1.5. Goals of the Present Work

This study focuses on evaluating the usability of BuddyGPT, an AI-powered chatbot. BuddyGPT is a prototype of an AI-powered chatbot that features a newly designed interaction strategy. Instead of offering direct answers, BuddyGPT is designed to support users by providing information and locating the relevant academic resources. For example, when asked about assignment answers, users can use the information from the BuddyGPT to solve the problem rather than directly getting the answers from BuddyGPT.

Our goal is to assess the current usability of the chatbot BuddyGPT and, based on the study's findings, provide recommendations to developers for improving the usability of BuddyGPT. Accordingly, this study takes an exploratory approach to investigate.

Frøkjær et al. (2000) found that there is only a weak correlation between satisfaction, effectiveness, and efficiency, but the application domain and context of use can lead these three factors have other complex relations. In addition, Baquero (2022) reports a strong positive correlation between satisfaction levels and individual-level Net Promoter Score (NPS), indicating that increases in satisfaction yield higher individual-level NPS. Net Promoter Score (NPS) is defined as a kind of indicator to measure customers' or users' willingness to recommend a product or service and their loyalty (Mandal, 2014).

According to Billestrup et al. (2016), the gender and educational background of users do not correlate with satisfaction. So, we plan to verify that the users' gender and educational background are not correlated with satisfaction. Moreover, we want to verify that satisfaction would not be affected by the previous usage frequency of chatbots in line with Lee (2000). Based on these studies, we pose the first research question (RQ1):

- RQ1: Is satisfaction with BuddyGPT affected by individual characteristics of the respondents, e.g., gender, educational background, and previous usage frequency of chatbots?

According to the definition of usability (ISO 9241-11, 2019) and the study from Santa et al. (2019), the level of satisfaction in using BuddyGPT is positively affected by effectiveness and people's efficiency. Building on these findings, we propose the second research question (RQ2):

- RQ2: Is users' satisfaction rate significantly affected by their performance in using BuddyGPT, e.g., effectiveness and people efficiency?
    - People's effectiveness in course-related information retrieval using the chatbot is measured as the number of tasks achieved and the number of reformulations of queries (task success rate and query reformulation rate).
    - People's efficiency in using the chatbot, measured by the time spent completing tasks.

According to the study of Baquero (2022), satisfaction has a positive influence on the recommendation intention of users. However, other scholars have suggested that satisfaction and individual-level NPS are only weakly correlated (Leslie et al., 2022). Based on this, we formulate the third research question (RQ3)

- RQ3: Is there a correlation between satisfaction declared by participants and their intention to recommend BuddyGPT, measured by individual-level Net Promoter Score (NPS) as rated by participants?

Prior research suggests that a system with high levels of effectiveness, efficiency, and satisfaction often achieves a high individual-level Net Promoter Score (NPS) (Pradini et al.,2019). Thus, we investigate how users' efficiency, effectiveness, and satisfaction affect individual-level NPS of BuddyGPT.

- RQ4: Are individual-level Net Promoter Scores (NPS) of BuddyGPT affected by users' efficiency, effectiveness, and satisfaction?

2. Methods

2.1. Design

This study adopted a within-subjects experimental design, integrating a standardized online survey (Qualtrics) with task-based interaction in a controlled laboratory setting. During the experiment, two laptops were used: one to run the BuddyGPT program, and the other for participants to complete the questionnaire. Ethical approval was obtained from the BMS Ethics Committee at the University of Twente, and all participants signed informed consent before participation (Appendix A).

The survey instrument was hosted on Qualtrics and comprised one informed consent of the experiment and three phases, with a pre-interaction phase, interaction phase, and post-interaction phase. Before participants filled the questionnaire, participants were provided informed consent and reviewed an introduction that outlined the study's purpose, procedures, withdrawal rights, data protection measures, and contact information; they could proceed only after clicking "I Agree."

During the pre-interaction phase, a questionnaire—divided into two sections—was administered to collect comprehensive background information and prior experience with chatbots:

- Section 1: Demographic Information – Participants provided basic details such as age, gender, and academic year (see Appendix B).
- Section 2: Previous Experience with Chatbots – This section mainly explored participants' prior exposure to chatbots, such as frequency of use, their perceived skill and confidence in using AI chatbots for learning, and some open-ended questions about participants' expectations of academic chatbots (see Appendix B).

In the interaction phase, each participant was presented with three tasks. Each task was structured with a task name, background description, and task instructions. Participants were instructed to formulate their inquiries based on the provided task instructions and then use these self-generated questions to interact with BuddyGPT (see Appendix E).

Task 3 was more complicated than tasks 1 and 2 in the design. But participants were not informed about the difficulty levels of the tasks, and these three tasks were presented in a randomized order. They were responsible for generating their query sentences to interact with BuddyGPT, completing the given tasks, and subsequently rating each task's difficulty. Throughout the experiment, all query dialogues were recorded as text, and the interaction time (measured in seconds) was captured from the start of the interaction until its conclusion.

In the post-interaction phase, participants completed the BUS-11 questionnaire and rated their likelihood of recommending BuddyGPT, answering NPS.

Quantitative data were collected throughout the entire process. In the pre-interaction phase, participants' previous AI usage was documented via a multiple-choice questionnaire. During the interaction phase, responses to control questions were collected using multiple-choice options, and task difficulty was assessed using a seven-point Likert scale (ranging from "extremely difficult" to "extremely easy"). Finally, in the post-interaction phase, the likelihood of recommending BuddyGPT was measured using the Net Promoter Score (0–10), while responses on the BUS-11 scale were captured via a five-point Likert scale (from "strongly disagree" to "strongly agree").

2.2. Participants

This study employed convenience sampling (N = 53) to recruit participants from the BMS Test Subject Pool Website (University of Twente, 2025). The resulting sample primarily consisted of Bachelor's, Master's, and PhD students from various programs at the University of Twente, with no restrictions regarding study fields, sex, nationality, or age. The diverse educational background of participants is close to the real users of the BuddyGPT and ensures the representativeness of the study.

The study was approved by the Ethics Committee of the Faculty of Behavioral, Management, and Social Sciences (BMS) at the University of Twente, and participants received course credits for their involvement via the Test Subject Pool Website (University of Twente, 2025).

2.3. Materials

The procedure of BuddyGPT was installed on a laptop computer running a Windows system, and all participants interacted with BuddyGPT through this device. Another laptop was used to run Qualtrics for participants to answer the questionnaire. And 1-minute tutorial video was prepared for participants to demonstrate the query process to familiarize them with how to use the system.

Before the experiment, participants were provided with a consent form that detailed the study's purpose, procedures, rights of withdrawal, potential risks and benefits, and data confidentiality (see Appendix A). The materials also included a standardized satisfaction scale embedded within the survey and a single-item Net Promoter Score question.

2.4. Measures

According to ISO 9241-11 (2018), usability is defined as the extent to which a system can be used effectively, efficiently, and satisfactorily in a specific context. In this study, the satisfaction of the BuddyGPT was evaluated using the four-factor BUS-11 scale (Borsci et al., 2024). The most recent version of the BUS-11 scale adopts a four-factor structure, in which the previously separate "Conversation" and "Functionality" dimensions—due to their high intercorrelation—have been combined into a unified factor named "Functional Interactive Conversation." From psychometric and designometric perspectives, this change makes the BUS-11 more parsimonious and accurate. Empirical validation of the four-factor BUS-11 has demonstrated robust psychometric properties, with Alpha = 0.89. Moreover, this scale was chosen because it offers a concise yet comprehensive evaluation of chatbot satisfaction by integrating key dimensions into four distinct factors (see Appendix B). The first factor is accessibility, which evaluates the ease with which users can detect and locate chatbot functions, ensuring that essential features are readily available. The second factor, functional interactive conversation, is a merged dimension that integrates elements of both conversational quality and system functionality. It reflects the chatbot's ability to communicate clearly, maintain contextual coherence, and deliver information interactively in a way that supports users' goals. Privacy is the third factor, and this factor measures users' perceptions regarding the chatbot's communication about privacy matters, such as how it informs users about data usage and potential privacy issues during interactions. The final factor is responsiveness. This factor reflects the promptness of the chatbot's replies, assessing how quickly the system responds to user queries.

To measure the effectiveness of BuddyGPT, the task success rate was used (Shawar & Atwell, 2007). The task success rate evaluated the percentage of successfully completed tasks where participants were able to finish tasks by retrieving the correct course-related information using BuddyGPT. A control question was used to assess correctness. If participants responded correctly based on BuddyGPT responses, the task was marked as successful. In this study, we added a reference variable, query reformulation rate, to assess how often users need to rephrase or modify their queries to receive relevant responses. And the number of query reformulations excludes the initial query submitted to BuddyGPT. A high reformulation rate may indicate difficulties in understanding the chatbot's guiding responses, while a lower rate suggests that users can effectively retrieve information with minimal adjustments.

In order to measure the efficiency of the chatbot, task completion time was recorded for each task. (Shawar & Atwell, 2007). It refers to the total time (in seconds) taken by participants to complete a given task using BuddyGPT, from the moment they receive the task instructions to the point of retrieving the required information. Moreover, task completion time provides insights into the system's speed and efficiency in facilitating information retrieval.

The likelihood of recommending BuddyGPT was measured using the individual-level Net Promoter Score. Individual-level NPS is a highly efficient scale to measure the satisfaction of users or a group (Sasmito et al., 2019). The question of individual-level NPS is commonly "how is the possibility of you to recommend the product to your friends or family," and the range of rating is 0 to 10. Participants are divided into three categories based on their rating result, with 0-6 points being Detractors, 7-8 points being Passives, and 9-10 points being Promoters. Individual-level NPS scores can be aggregated to derive a group-level NPS, reflecting the overall likelihood of the group to recommend the product or service. The calculating formation of group-level NPS is Promoter% - Detractor% = group-level NPS. The rating range of group-level NPS is -100 to 100.

- Group-level NPS > 0: More promoters than detractors, indicating a good brand reputation and overall positive user experience.
- Group-level NPS > 50: It is considered excellent, signifying high user loyalty and a strong willingness to recommend the brand.
- Group-level NPS < 0: More detractors than promoters, suggesting a low satisfaction and a poor user experience.

2.4. Procedure

Before the experiment began, the researcher provided a detailed introduction of the entire procedure to the participant. During the experiment, participants completed the questionnaires via Qualtrics, a web-based survey platform. First, participants reviewed and signed a consent form that explained the data collection policy. Next, they watched a guiding video demonstrating how to use BuddyGPT. Then, participants interacted with BuddyGPT to answer three multiple-choice questions based on specific tasks. During this stage, they also reported which task they were addressing and informed the researcher upon completing their interaction with BuddyGPT so that the researcher could accurately record the interaction time. Finally, participants completed a questionnaire assessing satisfaction, individual-level NPS, and the perceived advantages and deficiencies of BuddyGPT.

2.5. Data analysis

The data was cleaned by using Excel and analyzed using R from the following perspectives. Firstly, using Excel and R to do data preprocessing and descriptive statistics. All interaction durations recorded in the “MM:SS: ms” format (task completion time) were converted to seconds (s). Descriptive statistics were used to have a general overview of the data, which was done by using the R program to report the minimum, quartiles, median, mean, and maximum values of each variable.

Secondly, perform outlier removal and conduct distribution tests. For each continuous variable (difficulty of tasks, tasks, BUS-11 value, task completion time, and individual-level NPS), outliers were removed using the Interquartile Range Method to ensure the robustness of subsequent analyses. However, the final decision to remove outliers was based on whether their exclusion was deemed reasonable in the context of the analysis.

Thirdly, scale reliability, distributional assumption testing, and order effect testing. To ensure that the BUS-11 scale coherently measures the intended construct, internal consistency was calculated by computing Cronbach’s α for the full 11-item, four-factor instrument, thereby establishing a reliable foundation for subsequent correlation and regression analyses. Histograms and Q–Q plots were generated to visually inspect the distribution patterns. In the distribution tests, the Shapiro–Wilk normality test was used to assess whether difficulty of tasks and BUS-11 scores followed an approximately normal distribution. In the order effect testing, Levene’s test for homogeneity of variances was conducted to examine whether the variances were equal across the six randomly assigned presentation order groups (hereafter referred to as the task order group) (see Appendix F). The factor tasks order group needed to be examined for its potential effect on the variables, with difficulty of tasks, BUS-11 scores, and individual-level NPS. So, under the assumptions of normality and homogeneity of variances, separate one-way ANOVAs were conducted on difficulty of tasks, BUS-11 scores, and individual-level NPS to compare mean scores across the six task order group sequences.

In order to answer our research question, we perform correlation and regression analyses. Compute pairwise Pearson correlation coefficients among the key variables by referring to the research questions. Specifically, to answer the first research question, we examined whether the users’ satisfaction with BuddyGPT will be affected by demographic and individual aspects of the respondents. So, the potential significant associations between participants’ gender, educational background, and their satisfaction with BuddyGPT will be statistically assessed. And the correlation between prior use frequency and participants’ satisfaction will be tested.

The second research question focuses on examining the relationship between effectiveness (measured by task success rate and query reformulation rate), efficiency (measured by task completion time), and satisfaction. To address this, correlations between effectiveness, efficiency, and satisfaction will first be assessed. If significant correlations are found, a regression analysis will be conducted, with effectiveness and efficiency as independent variables and satisfaction as the dependent variable.

The third research question investigates whether there is a correlation between individual-level NPS and satisfaction. If a significant relationship is observed, a linear regression will be performed with individual-level NPS as the dependent variable and satisfaction as the independent variable, accompanied by a regression line plot.

The fourth research question investigates whether effectiveness, efficiency, and satisfaction can predict changes in individual-level NPS. Pearson correlation tests will first be conducted to assess the relationships between each variable and individual-level NPS. Multicollinearity among the three predictors will then be evaluated, and if present, variables may be combined. If significant correlations exist, a multiple linear regression will be performed to assess the collective predictive impact of these usability factors on individual-level NPS.

3. Result

3.1. Descriptive Analysis of Participants, Chatbot Use, and Task Performance

The sample consisted of 53 participants (male = 27, female = 26). However, chat history for four participants was lost due to a technical issue, so all subsequent analyses are conducted on the remaining 49 participants (male = 24, female = 25). Participants range in age from 19 to 33 years (*M* = 23.5, *SD* = 3.2). The majority fall into the 18–24 age group (*n* = 37, 69.8%), followed by 25–30 (*n* = 14, 26.4%) and 31–35 (*n* = 2, 3.8%). Regarding educational background (see Figure 1), most participants are third-year bachelor's students (*n* = 24), followed by PhD students (*n* = 11).

Figure 1

The Distribution of Educational Background of Participants

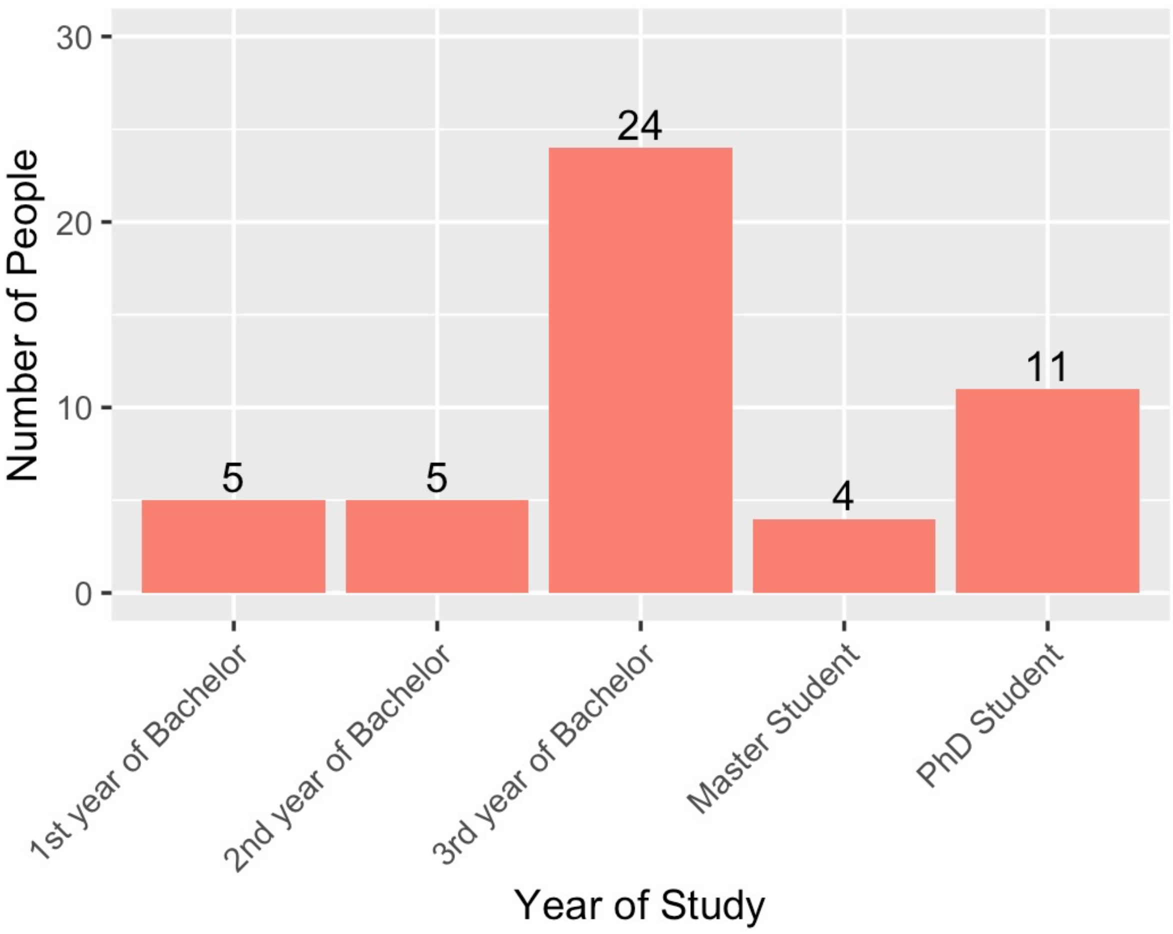


Figure 2 shows that the majority of participants use AI-powered chatbots frequently for academic purposes, with nearly 78% reporting regular use (daily or more often). The most common frequency is “almost daily” (21 participants), followed by “daily” and “once a week.” In contrast, only a small minority use them infrequently, such as monthly or less.

Figure 2

Usage Frequency of Various Chatbot Platforms

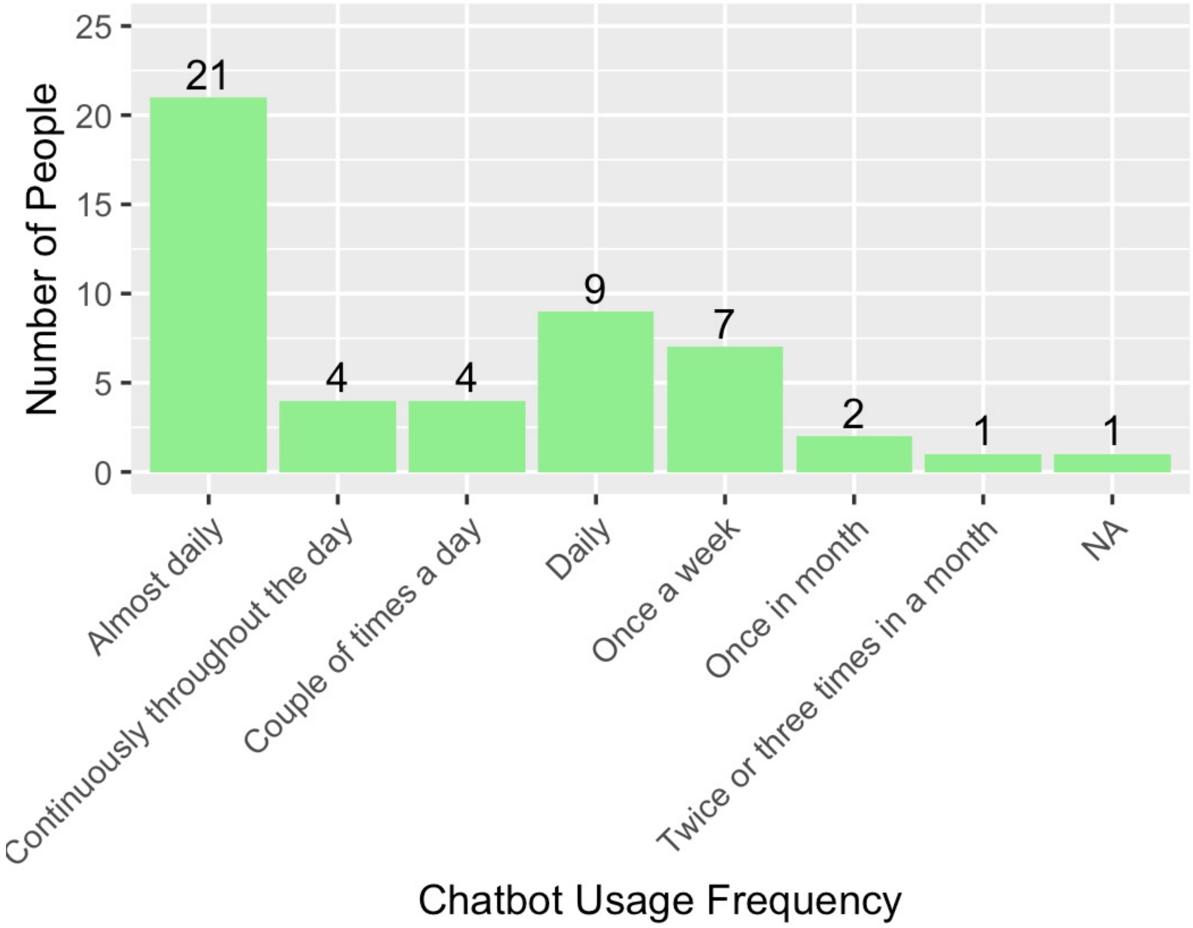


Figure 3 illustrates a strong preference for ChatGPT among participants, with minimal adoption of alternative chatbots. While a small number also explore tools like DeepSeek, Google Bard (Gemini), and Microsoft Copilot, the data suggest that ChatGPT overwhelmingly dominates academic chatbot usage in this sample.

Figure 3

Frequency Distribution of Various Chatbot Platforms Used for Academic Purposes

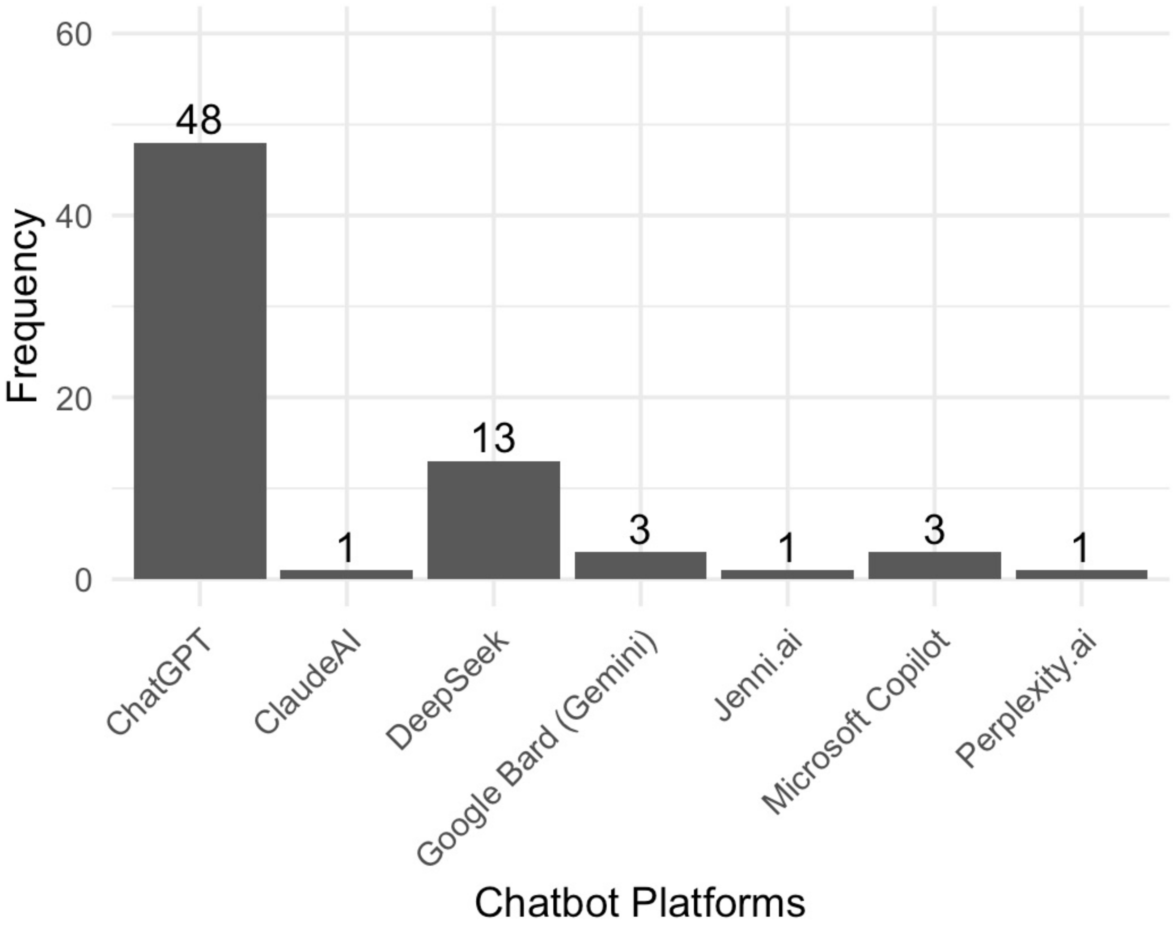


Table 1 presents the order groups. Participants are allocated unevenly across the six task order groups: orders 4 and 1 comprise the largest cohorts (14 and 10 participants, respectively), orders 2, 5, and 6 each account for moderate-sized cohorts (7 participants apiece), and order 3 encompasses the smallest cohort (4 participants). As the number of participants in each order group varies, it is necessary to examine whether the order group is an independent variable influencing the dependent variables, such as satisfaction and individual-level NPS.

Table 1

Distribution of Participants Across Task Order Groups and Task Sequences

Group	Task 1	Task 2	Task 3	Number
1	1	2	3	10
2	3	2	1	7
3	2	3	1	4
4	2	1	3	14
5	3	1	2	7
6	1	3	2	7

*Note***.** The numbers in the Task 1, Task 2, and Task 3 columns represent the order in which each task is presented to participants. For example, a value of 1 indicates that the task was shown first.

Table 2 presents participants' performance on tasks and perceived difficulty for tasks, including difficulty of tasks, task completion time, and success rate.

Table 2

Participants' Performance on Tasks

Tasks	Difficulty of Tasks (0-100)	Task Completion Time (s), M (SD)	Success Rate
Task 1	37.00	247.33 (108.63)	79.59%
Task 2	37.00	242.30 (115.56)	63.27%
Task 3	49.29	243.02 (123.73)	51.02%

Note. M = mean; SD = standard deviation; s = seconds.

Participants rate the third task as more challenging than the first and second tasks, with a perceived difficulty score of 49.29, compared to 37.00 for the other two. Task completion times are comparable across the three tasks, with mean durations of approximately 247.33

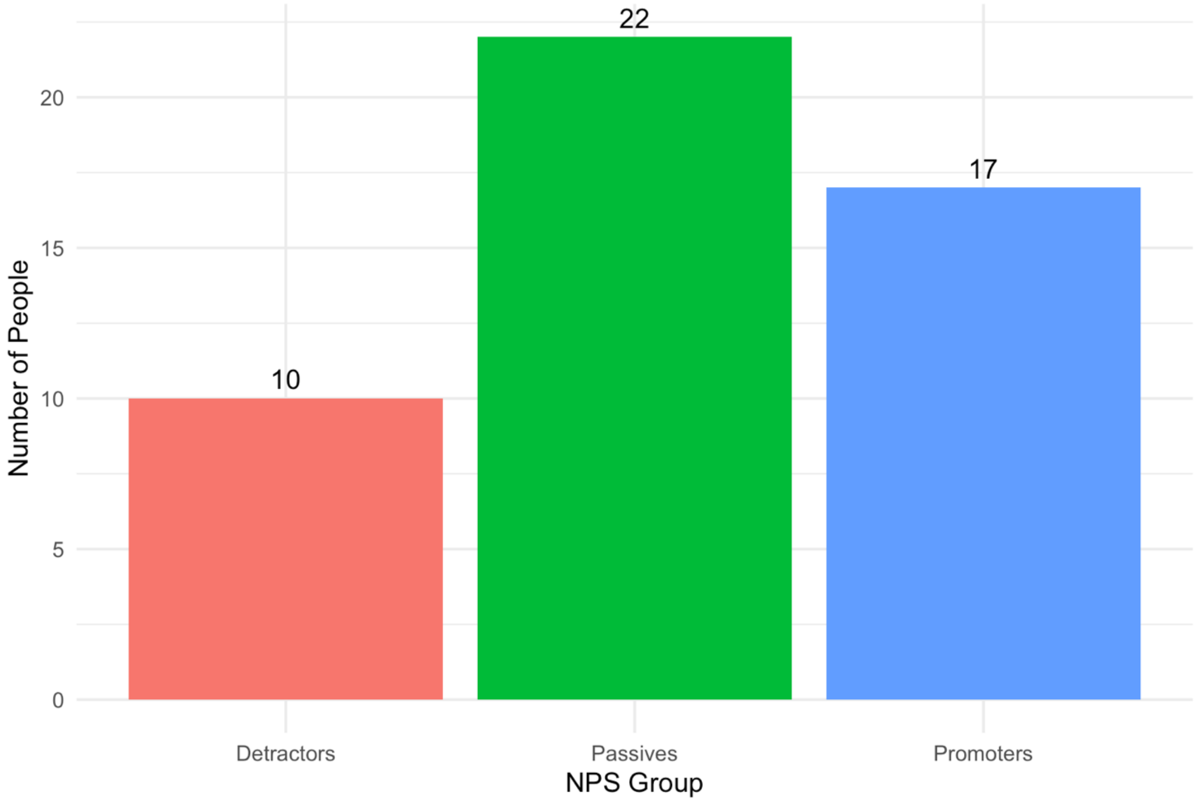
seconds for Task 1, 242.30 seconds for Task 2, and 243.02 seconds for Task 3. For the success rate of tasks, task 1 has the highest success rate (79.59%), task 2's success rate (63.27%) is close behind, and task 3 has the lowest success rate (51.02%). In addition, the mean of the satisfaction score is M = 3.45 (SD = 0.62) on a 5-point scale. It indicates that participants are, on average, satisfied with BuddyGPT. Approximately 68% of scores fall within one standard deviation of the mean (range: 2.82 – 4.06), indicating low variability in individual satisfaction ratings.

In correlational analysis, no significant correlation is found between task success rate and task completion time ($r = .03$, $p = .83$, 95% *CI* [–.25, .31]), indicating that participants' ability to complete tasks is not meaningfully associated with the amount of time they spend on them. Similarly, the correlation between task success rate and perceived task difficulty is not statistically significant ($r = –.19$, $p = .18$, 95% *CI* [–.45, .09]), suggesting that task success rate is not significantly influenced by how difficult participants perceive the tasks to be.

Regarding the individual-level NPS results, participants are categorized into three groups: Detractors (scores 0–6), Passives (scores 7–8), and Promoters (scores 9–10). The Passives group constitutes the largest proportion at 44.9%, followed by Promoters (34.7%) and Detractors (20.4%) (see Figure 4).

Figure 4

The Frequency Distribution of Individual-Level NPS Group (Detractors, Passives, Promoters)

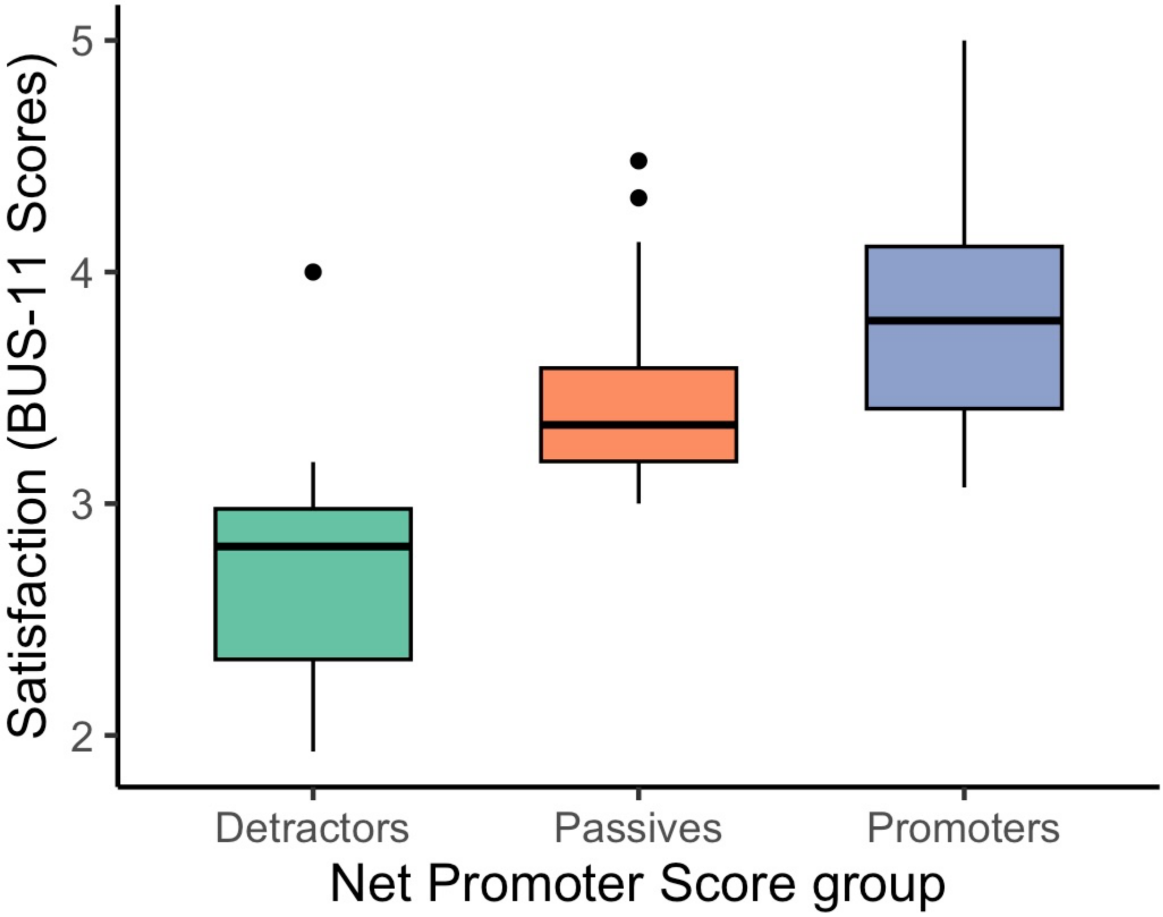


The group-level Net Promoter Score (NPS) is 14.3, which means that BuddyGPT has a good performance in the aspect of NPS. And the mean individual-level NPS of Detractors, Passives, and Promoters are 4, 7.5, and 9.19, respectively. These mean values fall near the center of their respective recommendation score ranges, rather than at the extremes.

Overall, satisfaction scores show a clear upward trend across individual-level NPS classifications, from Detractors to Passives to Promoters (see Figure 5). Specifically, Promoters' satisfaction scores are higher than those of the other two groups. And the median satisfaction score for Detractors is also close to 3, indicating that even Detractors feel somewhat satisfied with BuddyGPT.

Figure 5

The Distribution of Satisfaction Scores Among Detractors, Passives, and Promoters



3.2. Outlier treatment and distribution tests

For difficulty of tasks (*M* = 41.14), observation ID 28 and ID 33 are outliers, with values of 76.14 and 81.00, respectively. But the outlier ID 28 and ID 33 are retained because they both have a low task success rate. So, it is reasonable for them to give high value to the difficulty of tasks. For score of BUS-11(*M* = 3.45), observations ID 8 and ID 10 are outliers, but they are both retained, with the values of 5.00 and 1.93, respectively. Observations ID 8 and ID 10 could be reasonably explained by their number of successful tasks (3 and 2, respectively) and corresponding individual-level NPS ratings (9 and 2, respectively). For the individual-level NPS (*M* = 7.39), observations ID 2, 10, 31, and 41 are outliers, with individual-level NPS are 0, 2, 2, and 3, respectively. But they are all retained because these participants have a low task success rate and give low scores in the satisfaction survey. So, their low individual-level NPS is regarded as a resolvable situation.

3.3. Distributional Assumption Testing, Scale Reliability, and Task Order Effects

Based on Figures 6 and 7, the distributions of both satisfaction and difficulty of task scores align closely with the diagonal line, suggesting a visual approximation to normality. The BUS-11 usability scale yields a Cronbach's α of 0.79, which approaches the 0.80 ≤ α < 0.90 interval and thus indicates good internal consistency.

Figure 6

The Distribution of BUS-11 Scales

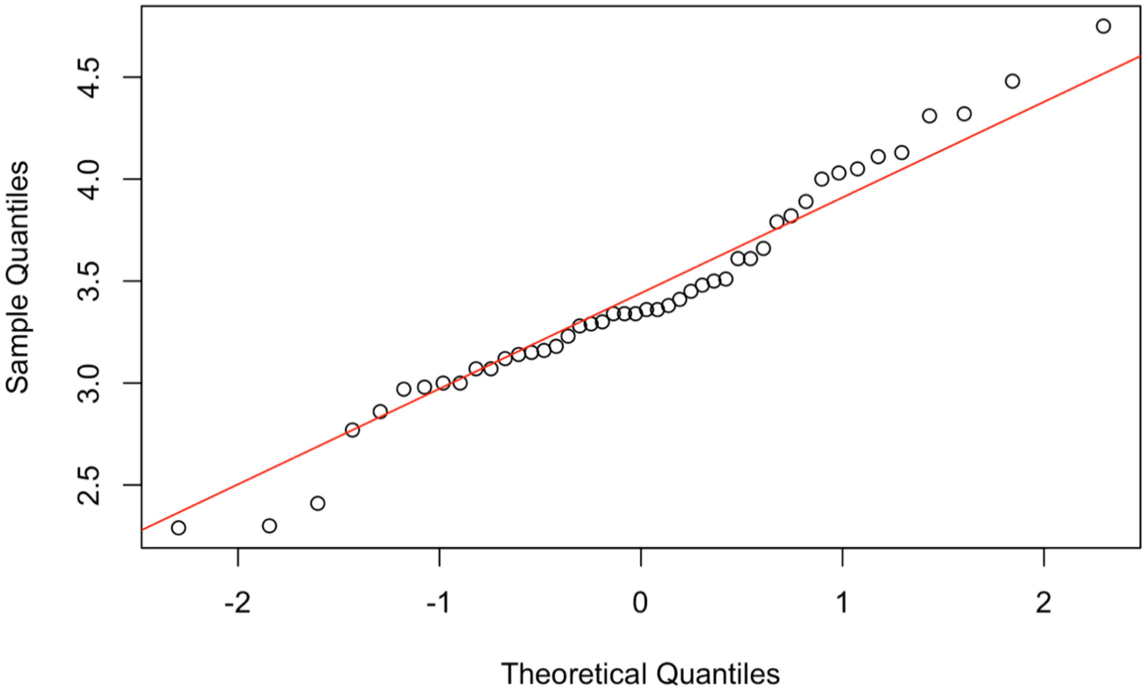
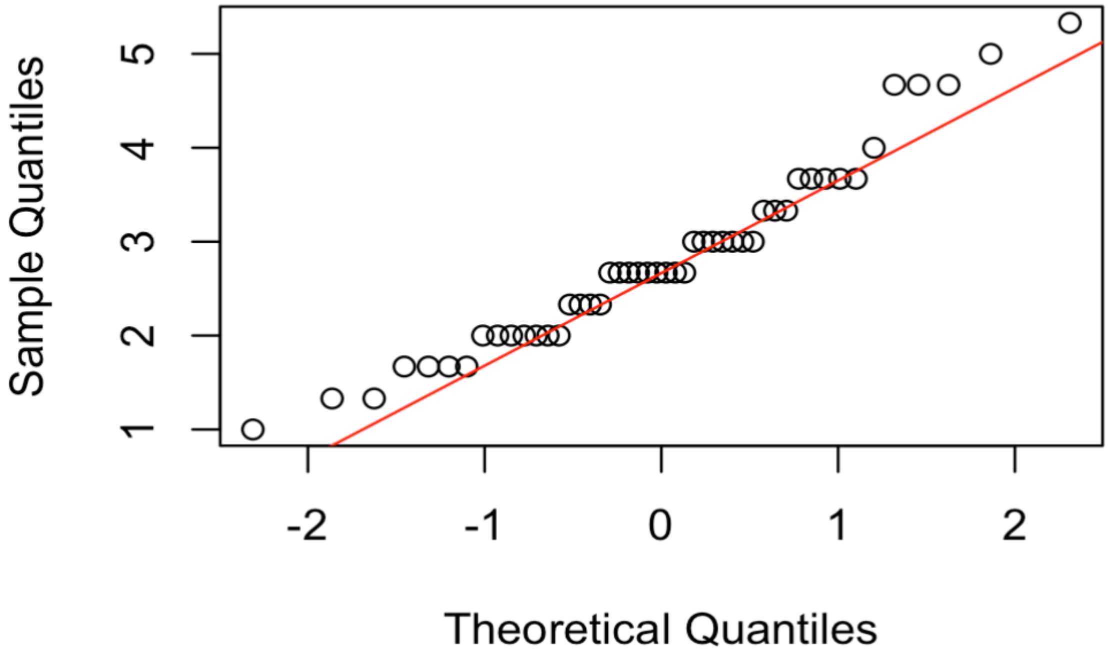


Figure 7

The Distribution of Difficulty of Tasks



The normal distributions of satisfaction and difficulty of tasks scores are also supported by the Shapiro–Wilk tests for normality, which yielded *W* = .96, *p* = .066 for difficulty of tasks and *W* = .97, *p* = .358 for satisfaction of BuddyGPT. As both p-values exceed the 0.05 threshold, the data do not significantly deviate from normality, allowing the distributions of satisfaction and difficulty to be treated as approximately normal (Shapiro & Wilk, 1965).

Additionally, Levene's tests confirm the assumption of homogeneity of variances, with *F*(5,42) = .60, *p* = .704 for difficulty of tasks, and *F*(5,40) = 1.56, *p* = .195 for satisfaction. Since both p-values exceed the conventional threshold of 0.05 (Shapiro & Wilk, 1965), the results indicate that variances across the six task order groups are not significantly different, thus supporting the assumption of homogeneity. After confirming the assumptions of normality and homogeneity of variances, one-way ANOVAs are conducted to examine the effect of task order group on each dependent variable. For difficulty of tasks, the ANOVA result is *F*(5,42) = .71, *p* = .621. For satisfaction, the result is *F*(5,40) = 1.74, *p* = .147. These results indicate that the presentation order of tasks had no significant effect on either perceived task difficulty or satisfaction scores.

ANOVA results revealed no significant group differences, suggesting that the random presentation order does not systematically influence participants' difficulty ratings or satisfaction scores, and thus ruling out order bias.

3.4. Participants' Performance Affects Satisfaction for BuddyGPT

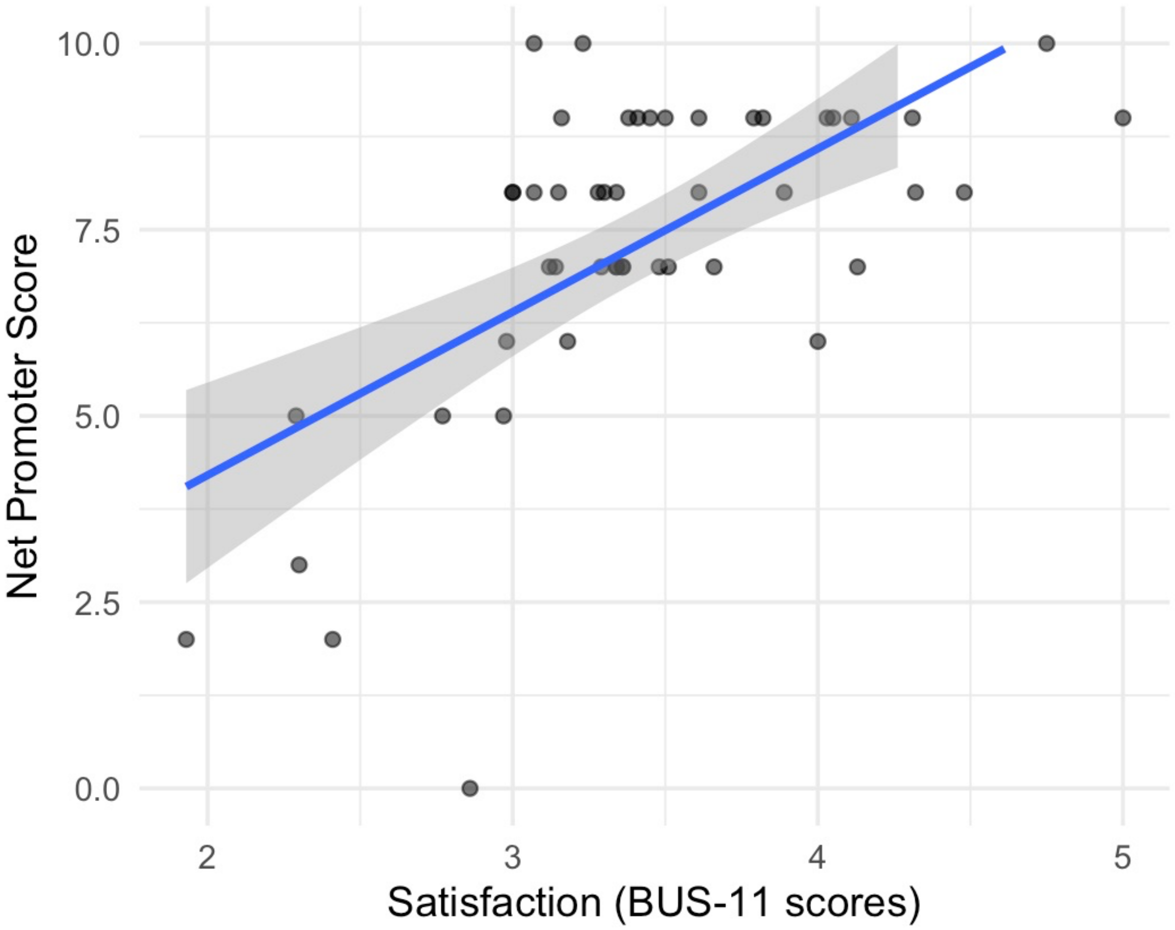
For the first research question, an independent samples *t*-test is conducted to compare the mean satisfaction scores between male and female participants. The results indicate no significant difference between the two groups, *t*(46.90) = –.50, *p* = .62. A linear regression analysis examines whether educational background predicts satisfaction with BuddyGPT. The model is not statistically significant, *F*(4, 44) = 1.61, *p* = .19, indicating that educational background does not significantly predict satisfaction scores. Similarly, a linear regression analysis is performed to evaluate whether prior usage frequency predicts participants' satisfaction with BuddyGPT. The result is also not significant, *F*(6, 41) = 1.13, *p* = .36, suggesting that prior usage frequency is not a significant predictor of satisfaction.

For the second research question, satisfaction is not significantly associated with task success rate (*r* = .09, *p* = .548, 95% *CI* [–.20, .36]) or number of reformulations (*r* = –.07, *p* = .642, 95% *CI* [–.33, .20]), indicating that participants' satisfaction with BuddyGPT is not related to effectiveness, as measured by task success rate and query reformulation rate. Similarly, satisfaction is not significantly correlated with task completion time (*r* = .13, *p* = .384, 95% *CI* [–.16, .40]), suggesting no association with efficiency, as measured by task completion time. Given the lack of significant correlations, further regression analysis is not deemed necessary.

For the third research question and the final research question, Pearson correlation tests are conducted to examine whether effectiveness, efficiency, and satisfaction are significantly associated with individual-level NPS scores. The results show that task success rate is not significantly correlated with individual-level NPS, *r* = .11, *p* = .45, 95% *CI* [–.18, .38], nor is query reformulation rate, *r* = –.09, *p* = .52, 95% *CI* [–.35, .19], indicating that effectiveness is not significantly related to individual-level NPS. Similarly, task completion time (efficiency) is also not significantly correlated with individual-level NPS, *r* = .20, *p* = .18, 95% *CI* [–.09, .44]. However, individual-level NPS exhibits a strong positive correlation with satisfaction (*r* = .63, *p* < .001, 95% *CI* [.44, .76]), demonstrating that better satisfaction perceptions are closely linked to greater intention to recommend the system (Figure 8).

Figure 8

Scatter Plot with Regression Line of Individual-Level NPS by Satisfaction Score

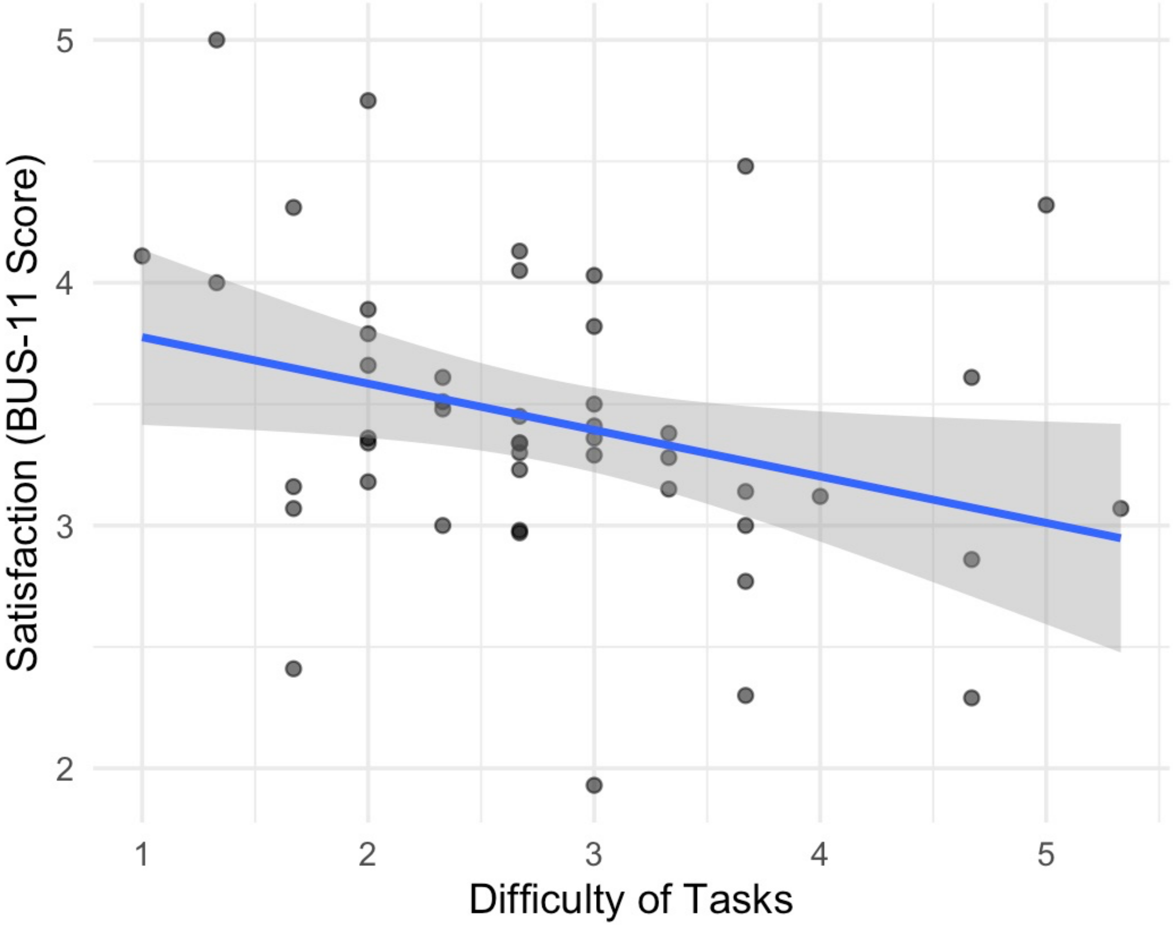


Since only satisfaction is significantly associated with individual-level NPS, multicollinearity among predictors is not a concern, and a simple linear regression using satisfaction as the sole predictor of individual-level NPS is sufficient. According to the result of linear regression, the intercept of the model is –.29 (*SE* = 1.36), but this value is not statistically significant, *t* (47) = –.21, *p* = .83, and can therefore be interpreted as negligible. The slope coefficient for satisfaction is 2.23 (*SE* = 0.39), indicating that for every one-point increase in satisfaction, the individual-level NPS score increases by approximately 2.23 points (see Figure 8).

Additionally, we examine the relationships between perceived task difficulty and satisfaction, as well as between query reformulation rate and perceived task difficulty. Satisfaction exhibits a moderate negative correlation with perceived task difficulty (*r* = –.31, *p* = .033, 95% *CI* [-.53, -.03]), indicating that participants who perceived the tasks as more difficult tended to report lower satisfaction ratings (see Figure 9).

Figure 9

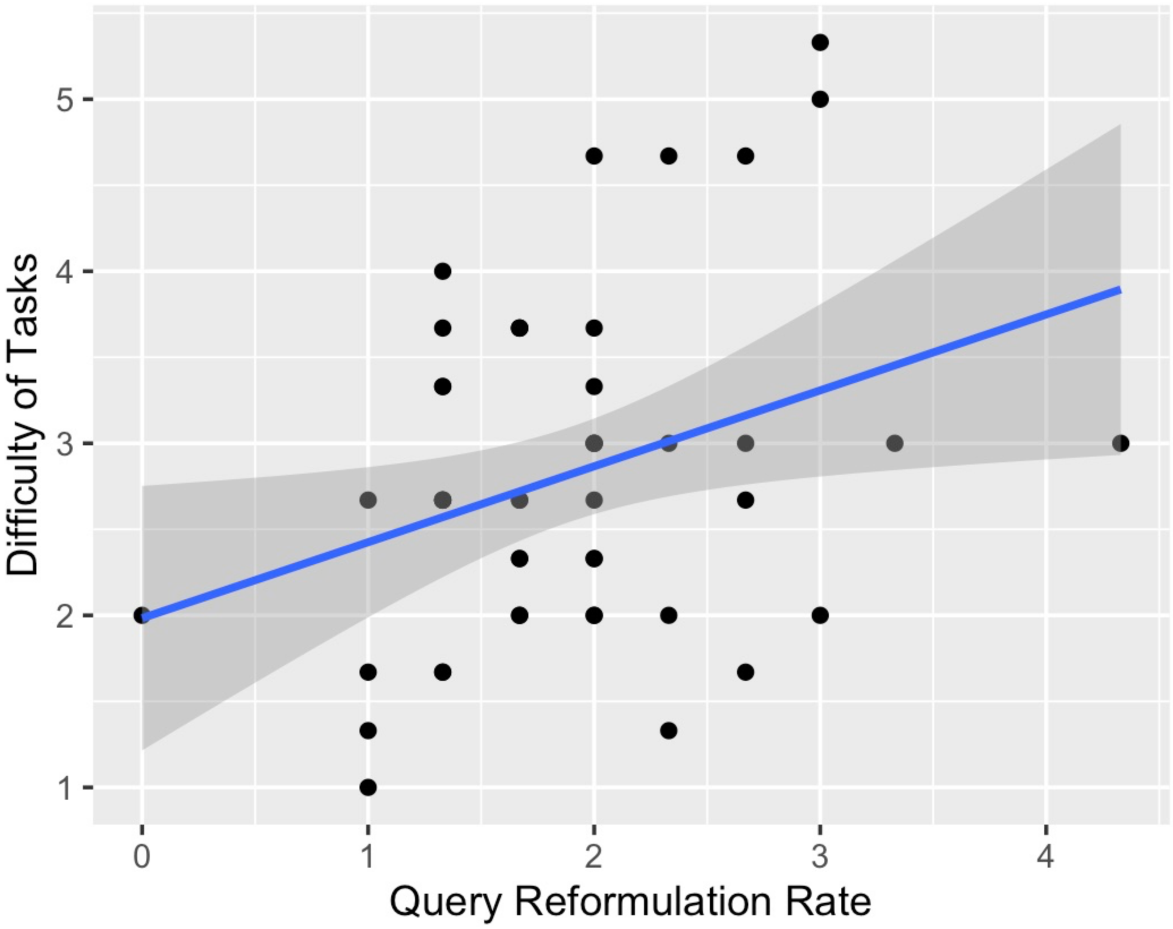
Scatter Plot with Regression Line of Satisfaction Score by Difficulty of Tasks



Moreover, perceived task difficulty is moderately positively correlated with the query reformulation rate (*r* = .33, *p* = .023, 95% *CI* [.04, .56]), suggesting that higher difficulty ratings are associated with more frequent reformulations (see Figure 10).

Figure 10

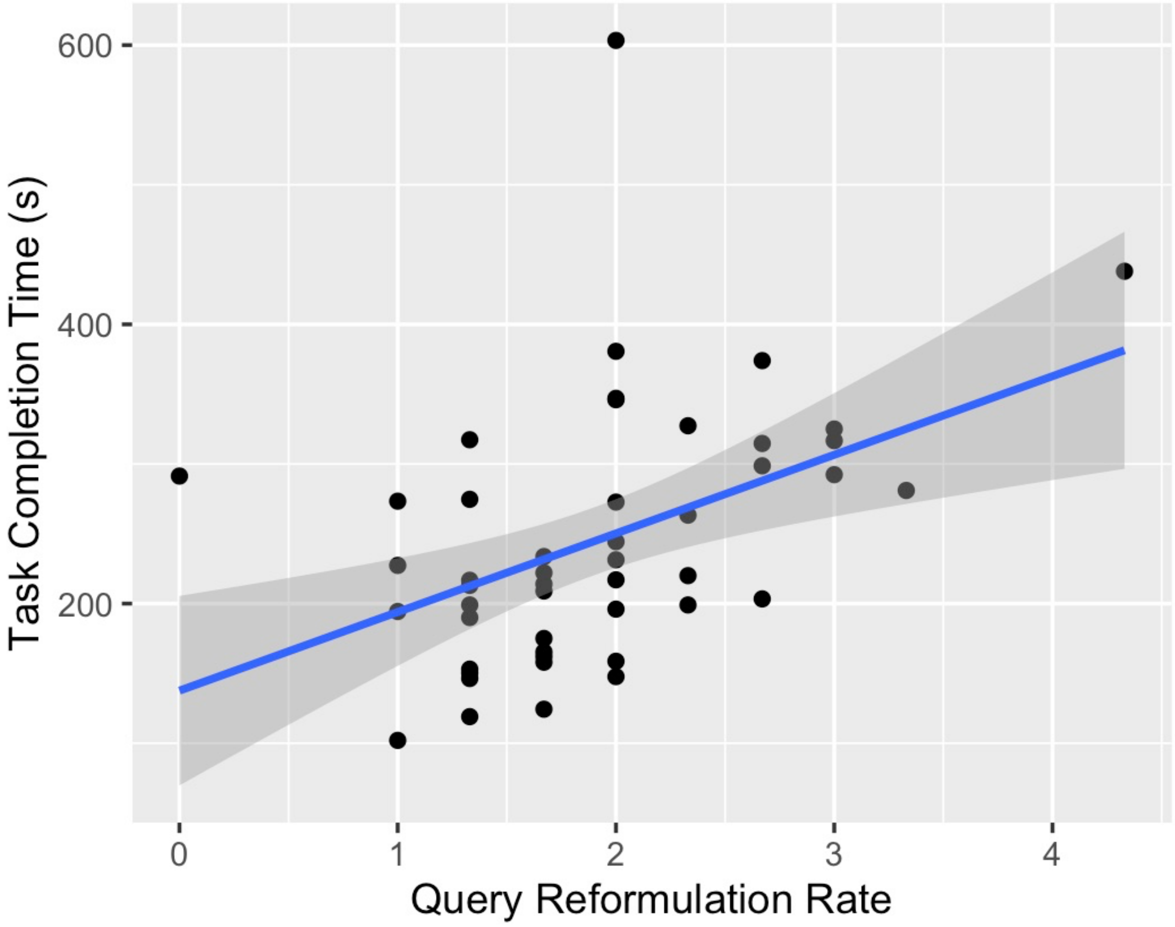
Scatter Plot with Regression Line of Difficulty of Tasks by Query Reformulation Rate



In addition, task completion time and query reformulation rate are also moderately positively correlated (*r* = .45, *p* = .001, 95% *CI* [.19, .65]), indicating that each additional reformulation adds approximately 56 seconds to the overall task duration (see Figure 11).

Figure 11

Scatter Plot with Regression Line of Task Completion Time by Query Reformulation Rate



4. Discussion

This study aimed to evaluate the usability of BuddyGPT, an AI-powered chatbot designed to assist university students in accessing course-related information. The study focused on three key usability dimensions – effectiveness, efficiency, and satisfaction – and further explored the relationship between these factors and users' recommendation intention, as measured by individual-level Net Promoter Score (NPS). The findings suggested that BuddyGPT demonstrated strong performance in helping users complete academic tasks, showing particularly high effectiveness in retrieving information from Canvas. And the result also indicated a strong correlation between satisfaction and the recommendation intent of users. These insights are particularly valuable for improving future chatbot design in education, highlighting the need to optimize not just functionality but perceived satisfaction among students.

4.1. Usability Insights Based on Participant Characteristics and Task Outcomes

Based on the results of descriptive statistics, we found that over 77.7% of participants had almost always used AI-powered chatbots to support their academic study, which meant majority of participants have rich experience in interacting with chatbots. So, we could eliminate the effect of unfamiliar chatbot usage skills on the assessment of BuddyGPT's usability. In addition, nearly all participants also claimed that they had used ChatGPT for academic purposes. Therefore, it could be inferred that participants' satisfaction ratings for the AI-powered chatbot BuddyGPT were likely made with ChatGPT as a reference point.

Based on participants' task performance, BuddyGPT indicated a strong ability to support students in retrieving the necessary academic information to finish their goals. And the low query reformulation rate indicated that BuddyGPT's responses were relatively accurate and understandable. But the average completion times across all three tasks were remarkably too long. Participants reported that the responses from BuddyGPT were overly verbose and difficult to find key information. And we found that BuddyGPT showed a lower success rate on more complex tasks, such as assisting students in completing their assignments. Based on participant feedback, we believe that Task 3's lower success rate and greater perceived difficulty resulted from overly verbose responses that distracted participants and prolonged their reading time. Therefore, we suggest that the developer of BuddyGPT should try to reduce the verbose and distracting information in BuddyGPT's responses, which can improve the efficiency and effectiveness of BuddyGPT. Specifically, developers could bold or highlight key information and have BuddyGPT generate more concise summaries to help users quickly locate essential details.

4.2. Factors Influencing Satisfaction and Their Correlation with Individual-Level NPS

The first research question aimed to examine the relationship between respondents' demographic and individual characteristics and their satisfaction with BuddyGPT. The results showed that gender, educational background, and prior chatbot usage experience were not significantly associated with satisfaction levels. Therefore, we conclude that users' satisfaction with BuddyGPT is independent of these factors—namely, gender, educational background, and prior chatbot usage experience—which aligns with the conclusion of Billestrup et al. (2016).

The second research question aimed to examine the correlation between users' satisfaction and the effectiveness and efficiency of BuddyGPT. The results showed that satisfaction is not significantly related to task success rate or query reformulation rate, which are the two indicators used to measure effectiveness. Therefore, we conclude that users' satisfaction with BuddyGPT is not associated with its effectiveness. Similarly, satisfaction was not significantly correlated with efficiency, as measured by task completion time, indicating that the time spent on tasks did not affect users' satisfaction with the system. This result challenges the integrated model of usability proposed by ISO 9241-11 (2019) and Santa et al. (2019), which conceptualizes effectiveness, efficiency, and satisfaction as interrelated components of usability. Similar findings have been reported by Frøkjær et al. (2000), who noted that these three dimensions may vary independently depending on context and user expectations. Thus, BuddyGPT enables effective and efficient task completion, but these aspects do not necessarily translate into satisfaction.

The third research question focused on the relationship between users' satisfaction with BuddyGPT and their intention to recommend it. And this relationship was studied by assessing the correlation between satisfaction and individual-level Net Promoter Score (NPS). The results revealed a strong positive correlation between satisfaction and the individual-level NPS, indicating that higher levels of satisfaction were associated with a greater likelihood of recommending the system. This finding supports the study of Baquero (2022) and challenges the study of Leslie et al. (2022). Therefore, we conclude that users' intent to recommend BuddyGPT is strongly influenced by their satisfaction with the system. Developers should use the four factors of the BUS-11 scale (Accessibility, Functional Interactive Conversation, Privacy, and Responsiveness) to guide design modifications, which ought to significantly enhance user satisfaction and thereby increase recommendation intent.

The fourth research question aimed to investigate whether effectiveness, efficiency, and satisfaction could serve as predictors of individual-level NPS. Based on the findings from the third research question, we had already established that satisfaction can predict individual-level NPS due to its strong positive correlation. However, our analysis showed that neither task success rate nor query reformulation rate was significantly correlated with individual-level NPS, suggesting that effectiveness is not associated with users' likelihood of recommending BuddyGPT. Additionally, task completion time was not significantly correlated with satisfaction, indicating that efficiency is also unrelated to satisfaction in this context. These findings do not align with the conclusion of Pradini et al. (2019).

To summarize, in the future development of BuddyGPT, developers cannot improve satisfaction by enhancing the system's effectiveness and efficiency. Therefore, this also implies that satisfaction must be addressed independently, potentially through other aspects of design or user experience. Given the strong correlation between satisfaction and individual-level NPS, it is necessary that increasing satisfaction can significantly boost positive word-of-mouth through user referrals, thereby enhancing the overall impact and adoption of BuddyGPT. Therefore, BuddyGPT's developers can regard the BUS-11 scale's four factors as a guideline to improve BuddyGPT.

4.3. Future Research

Firstly, due to the relatively small sample size, the generalizability of the results is limited. Additionally, since the experiment was conducted solely within the University of Twente, the diversity of the sample was further constrained. And owing to the use convenience sampling method, the number of participants in each educational level group was unequal, which may have influenced the applicability of the findings across users with different educational backgrounds. Therefore, in future research, the researchers can increase the number of samples, use a more randomized sampling method across a broader population to experiment, and ensure an equal number of participants in each condition group. These methods can improve the applicability of the experiment's conclusion to the population.

Secondly, technical limitations led to an unequal number of participants assigned to each task presentation order. As a result, additional effort was required during the data analysis phase to examine whether the order of presenting tasks had any effect on other factors. Although the results indicated no such effect, this does not fully rule out the possibility that task order may influence satisfaction or other variables through indirect pathways. So, the researcher can ensure an equal number of participants assigned to each task presentation order in future research.

Thirdly, some participants had limited English proficiency, which may have influenced certain experimental outcomes, such as task completion time. For example, the participants who are from non-English-speaking countries and study in the first year of a bachelor used a longer time to finish the tasks than other participants. Thus, in the future, the research can get more specific analysis results by grouping the participants based on whether they are from English-speaking countries.

Additionally, researchers can conduct further studies to explore the factors influencing users' satisfaction with BuddyGPT in the future. The goal is to provide insights that help developers enhance satisfaction, which in turn may increase the individual-level Net Promoter Score (NPS) and promote organic, user-driven dissemination of the system.

Although this study found that system effectiveness was not significantly correlated with users' satisfaction, this result may be due to the limited sample size. A study by Al-Maskari and Sanderson (2010) found a significant correlation between system effectiveness and satisfaction. Therefore, future research could involve a larger participant pool to re-examine the relationship between effectiveness and satisfaction.

In addition, Sanny et al. (2020) found that the perceived usefulness of a chatbot is significantly associated with satisfaction. Thus, investigating the role of chatbot usefulness could be another valuable direction for understanding satisfaction in the context of BuddyGPT.

The findings of this study indicate that BuddyGPT can reliably assist students in retrieving course-related information, thereby reducing the time spent searching for answers. This offers concrete benefits for self-directed learning: students can sustain their study momentum and concentrate on higher-order tasks (such as analysis and synthesis) rather than mere information gathering. Moreover, given BuddyGPT's demonstrated ability to handle more challenging tasks, students can more easily solve and learn complex problems with its support.

4.4. Conclusion

This research investigated the usability of BuddyGPT, a chatbot designed and developed by the University of Twente, through three key dimensions: effectiveness, efficiency, and satisfaction. In addition, the study examined the correlations between Net

Promoter Score and these three usability factors. The findings provide valuable insights and practical guidance for the developers of BuddyGPT to enhance its usability. Notably, the study revealed that effectiveness and efficiency were not significantly correlated with satisfaction, which contrasts with findings from previous traditional studies.

Overall, this research offers a new perspective on the factors influencing satisfaction and contributes to the ongoing development of chatbots aimed at improving the user experience. It also highlights the importance of incorporating individual-level NPS and the four factors of BUS-11 as key design considerations. This contribution can facilitate the rapid and widespread adoption of new AI-powered chatbots in the market, especially in the context of the growing integration of chatbots in the field of education.

5. References

Ahmad, S., Umirzakova, S., Mujtaba, G., Amin, M. S., & Whangbo, T. (2023). Education 5.0: Requirements, enabling technologies, and future directions [Preprint]. *arXiv*. https://doi.org/10.48550/arxiv.2307.15846

Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, *61*(5), 859–868. https://doi.org/10.1002/asi.21300

Anfoud, S., & Alami Talbi, L. F. (2024). Education 5.0: Perspectives Croisées du Japon et du Zimbabwe. *Analele Universității Din Craiova, Seria Psihologie-Pedagogie/Annals of the University of Craiova, Series Psychology-Pedagogy*, *45*(2 Suppl.), 55–70. https://doi.org/10.52846/aucpp.2023.2suppl.05

Baabdullah, A. M., Alalwan, A. A., Algharabat, R. S., Metri, B., & Rana, N. P. (2022). Virtual agents and flow experience: An empirical examination of AI-powered Chatbots. *Technological Forecasting and Social Change*, *181*, 121772. https://doi.org/10.1016/j.techfore.2022.121772

Baquero, A. (2022). Net promoter score (NPS) and customer satisfaction: Relationship and efficient management. *Sustainability*, *14*(4), 2011. https://doi.org/10.3390/su14042011

Benotti, L., Martinez, M. C., & Schapachnik, F. (2018). A tool for introducing computer science with Automatic Formative Assessment. *IEEE Transactions on Learning Technologies*, *11*(2), 179–192. https://doi.org/10.1109/tlt.2017.2682084

Bhavya, B., Gautam, G., & Sumedha, M. (2021). Impact of the COVID-19 pandemic on education system. *EPRA International Journal of Environmental Economics, Commerce and Educational Management*, *8*(2) 6–8. https://doi.org/10.36713/epra6363

Billestrup, J., Bruun, A., & Stage, J. (2016). Usability problems experienced by different groups of skilled internet users: Gender, age, and background. *Lecture Notes in Computer Science*, *9856*, 45–55. https://doi.org/10.1007/978-3-319-44902-9_4

Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A., & van der Velde, F. (2022). A confirmatory factorial analysis of the chatbot usability scale: A multilanguage validation. *Personal and Ubiquitous Computing*, *27*(2), 317–330. https://doi.org/10.1007/s00779-022-01690-0

Borsci, S., & Schmettow, M. (2024). Re-examining the chatbot usability scale (BUS-11) to assess user experience with Customer Relationship Management Chatbots. *Personal and Ubiquitous Computing*, *28*(6), 1033–1044. https://doi.org/10.1007/s00779-024-01834-4

Chang, C., Hwang, G., & Gau, M. (2021). Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology*, *53*(1), 171–188. https://doi.org/10.1111/bjet.13158

Daniel, S. J. (2020). Education and the COVID-19 pandemic. *PROSPECTS*, *49*(1–2), 91–96. https://doi.org/10.1007/s11125-020-09464-3

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/332040.332455

Hongli, Z., & Wai Yie, L. (2024). Industry 5.0 and education 5.0: Transforming Vocational Education Through Intelligent Technology. *Journal of Innovation and Technology*, *2024*(1). https://doi.org/10.61453/joit.v2024no16

International Organization for Standardization. (2018). *ISO 9241-11:2018 Ergonomics of human-system interaction—Part 11: Usability: Definitions and concepts.* Geneva, Switzerland: ISO. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

International Organization for Standardization. (2019). *ISO 9241-210:2019 Ergonomics of* *human-system interaction—Part 210: Human-centred design for interactive systems.* Brussels, Belgium: CEN. https://www.iso.org/standard/77520.html

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2022). Interacting with educational Chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Lee, A. T. (2000). Web site usability, usefulness, and visit frequency. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(4), 404–407. https://doi.org/10.1177/154193120004400401

Leslie, H. H., Lee, H.-Y., Blouin, B., Kruk, M. E., & García, P. J. (2022). Evaluating patient-reported outcome measures in Peru: A cross-sectional study of satisfaction and net promoter score using the 2016 ensusalud survey. *BMJ Quality &amp; Safety*, *31*(8), 599–608. https://doi.org/10.1136/bmjqs-2021-014095

Mandal, P. C. (2014). Net promoter score: A conceptual analysis. *International Journal of Management Concepts and Philosophy*, *8*(4), 209. https://doi.org/10.1504/ijmcp.2014.066899

Megahed, N. A., Abdel-Kader, R. F., & Soliman, H. Y. (2022). Post-pandemic education strategy: Framework for Artificial Intelligence-empowered education in engineering (aied-eng) for lifelong learning. *Lecture Notes on Data Engineering and Communications Technologies*, *113*, 544–556. https://doi.org/10.1007/978-3-031-03918-8_45

Ng, D. T., Leung, J. K., Su, J., Ng, R. C., & Chu, S. K. (2023). Teachers' ai digital competencies and twenty-first century skills in the post-pandemic world. *Educational Technology Research and Development*, *71*(1), 137–161. https://doi.org/10.1007/s11423-023-10203-6

Plantak Vukovac, D., Horvat, A., & Čižmešija, A. (2021). Usability and user experience of a chat application with integrated educational chatbot functionalities. *Lecture Notes in Computer Science*, *12785*, 216–229. https://doi.org/10.1007/978-3-030-77943-6_14

Pradini, R. S., Kriswibowo, R., & Ramdani, F. (2019). Usability evaluation on the SIPR website uses the system usability scale and net promoter score. *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, 280–284. https://doi.org/10.1109/siet48054.2019.8986098

Sain, Z. H., Ayu, S. M., & Thelma, C. C. (2024). Exploring the CHATGPT era: Finding equilibrium between innovation and tradition in Education. *Middle East Research Journal of Humanities and Social Sciences*, *4*(04), 116–121. https://doi.org/10.36348/merjhss.2024.v04i04.001

Sanny, L., Susastra, A. C., Roberts, C., & Yusramdaleni, R. (2020). The analysis of customer satisfaction factors which influence chatbot acceptance in Indonesia. *Management Science Letters*, 1225–1232. https://doi.org/10.5267/j.msl.2019.11.036

Santa, R., MacDonald, J. B., & Ferrer, M. (2019). The role of trust in e-government effectiveness, operational effectiveness, and user satisfaction: Lessons from Saudi Arabia in E-g2b. *Government Information Quarterly*, *36*(1), 39–50. https://doi.org/10.1016/j.giq.2018.10.007

Sara, R. D., Sonia, M. R., Eva, J. G., & Judit, R. L. (2023). The potential of educational chatbots for the support and formative assessment of students. In *New Trends and Promising Directions in Modern Education:" AI in Education"*, 105-136. https://hdl.handle.net/20.500.14352/101695

Sasmito, G. W., Zulfiqar, L. O., & Nishom, M. (2019). Usability testing based on system usability scale and net promoter score. *2019 International Seminar on Research of*

Information Technology and Intelligent Systems (ISRITI), 540–545. https://doi.org/10.1109/isriti48646.2019.9034666

Sedrakyan, G., Borsci, S., van den Berg, S. M., van Hillegersberg, J., & Veldkamp, B. P. (2024). Design implications for next generation chatbots with education 5.0. *Lecture Notes in Educational Technology*, 1–12. https://doi.org/10.1007/978-981-97-3883-0_1

Selwyn, N. (2013). Rethinking education in the Digital age. *Digital Sociology*, 197–212. https://doi.org/10.1057/9781137297792_14

Shahidi Hamedani, S., Aslam, S., Mundher Oraibi, B. A., Wah, Y. B., & Shahidi Hamedani, S. (2024). Transitioning towards Tomorrow's workforce: Education 5.0 in the landscape of Society 5.0: A systematic literature review. *Education Sciences*, *14*(10), 1041. https://doi.org/10.3390/educsci14101041

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shawar, B. A., & Atwell, E. (2007). Different measurements metrics to evaluate a chatbot system. *Proceedings of the Workshop on Bridging the Gap Academic and Industrial Research in Dialog TechnologieNAACL-HLTHLT' 07*, 89–96. https://doi.org/10.3115/1556328.1556341

Sun, D., Boudouaia, A., Zhu, C., & Li, Y. (2024). Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study. *International Journal of Educational Technology in Higher Education*, *21*(1). https://doi.org/10.1186/s41239-024-00446-5

University of Twente. (2025). *Test Subject Pool BMS*. https://utwente.sona-systems.com

Xu, X., Lu, Y., Vogel-Heuser, B., & Wang, L. (2021). Industry 4.0 and industry 5.0—inception, conception and perception. *Journal of Manufacturing Systems*, *61*, 530–535. https://doi.org/10.1016/j.jmsy.2021.10.006

Yang, S., & Stansfield, K. (2022). AI chatbot for Educational Service Improvement in the post-pandemic ERA: A case study prototype for supporting Digital Reading List. *2022 13th International Conference on E-Education, E-Business, E-Management, and E-Learning (IC4E)*, 24–29. https://doi.org/10.1145/3514262.3514289

6. Appendices

Appendix A: The Consent of The Questionnaire

Welcome to the study

You are being invited to participate in a research study on **"Usability and Overall User Experience Evaluation of the New-Generation Educational Chatbot (BuddyGPT)."** You will interact with the chatbot,a computer program that communicates with people using natural language through text to answer questions and provide information (e.g. Example Chatbot) This study is conducted by Jiawei Li, Cosmin G. Ghiauru, Nese Baz Aktas from the **Faculty of Behavioural, Management, and Social Sciences at the University of Twente.**

Purpose of the Study:

The purpose of this research is to evaluate the usability and overall user experience of **BuddyGPT**, an AI-driven educational chatbot designed to help students retrieve course-related information. Your participation will contribute to understanding the chatbot's effectiveness, usability, and potential improvements.

Procedure & Duration

The study will take approximately **30 minutes** to complete. You will interact with the chatbot, complete assigned tasks, and answer pre- and post-interaction survey questions.

Voluntary Participation

Participation in this study is entirely **voluntary.** You may choose to withdraw at any time without any consequences. You may also skip any question you do not wish to answer.

Confidentiality & Data Protection

We are committed to protecting your privacy. Your responses will be **anonymized and securely stored.** The data will only be used for research purposes and will not contain any personally identifiable information. All collected data will be stored securely and handled in compliance with University of Twente's ethical guidelines.

Potential Risks and Benefits

There are no anticipated risks associated with participating in this study. While participation will not directly benefit you, your insights will contribute to the development and refinement of educational chatbots.

Contact Information

If you have any questions or concerns regarding this study, you may contact:

📩 Jiawei – j.li-13@student.utwente.nl
📩 Cosmin G. Ghiauru – c.ghiauru@student.utwente.nl
📩 Nese Baz Aktas – nese.baz@utwente.nl

Supervisors

📩 Gayane Sedrakyan – g.sedrakyan@utwente.nl
📩 Simone Borsci – s.borsci@utwente.nl

Consent By clicking **"I Agree,"** you confirm that: ✓ You have read and understood the study details. ✓ You are voluntarily participating in this research. ✓ You understand that you can withdraw at any time.

○ **I Agree** (1)

○ **I Do Not Agree** (2)

Appendix B: Excerpt of The Questionnaire of The Experiment

Note. The entire survey is available upon request to Dr Borsci

Demographics1 What is your age?

____

Demographics2 What is your sex?

○ Male (1)

○ Female (2)

○ Prefer not to say (3)

Demographics3 What is your Major?

____

Demographics4 What is your current year of the study in your academic program?

○ 1st year of Bachelor (1)

○ 2nd year of Bachelor (2)

○ 3rd year of Bachelor (3)

○ Master Student (4)

○ PhD Student (5)

Page Break

Prior Experience1 Have you ever used a chatbot? (Examples: ChatGPT, Microsoft Copilot, Duolingo Bot, customer service chatbots)

○ No (1)

○ Yes (2)

Display this question:

If Have you ever used a chatbot? (Examples: ChatGPT, Microsoft Copilot, Duolingo Bot, customer servi... = Yes

Prior Experience2 If you think about the last month, how many times have you used chatbots?

○ Never (1)

○ Once in month (2)

○ Twice or three times in a month (3)

○ Once a week (4)

○ Almost daily (5)

○ Daily (6)

○ Couple of times a day (8)

○ Continuously throughout the day (9)

Display this question:

If Have you ever used a chatbot? (Examples: ChatGPT, Microsoft Copilot, Duolingo Bot, customer servi... = Yes

Prior Experience3 Which chatbot(s) do you prefer to use for everyday tasks or personal use unrelated to university duties? (You may select multiple options.)

□ ChatGPT (1)

□ Google Bard (Gemini) (2)

□ Microsoft Copilot (3)

□ Meta AI (4)

□ Deepseek (5)

□ I am not using chatbots for non university related tasks (6)

□ Other (please specify—enter all that apply, separated by commas): (7)
____

Display this question:

If Have you ever used a chatbot? (Examples: ChatGPT, Microsoft Copilot, Duolingo Bot, customer servi... = Yes

If Have you ever used a chatbot? (Examples: ChatGPT, Microsoft Copilot, Duolingo Bot, customer servi... = Yes

Prior Experience5 Which chatbot/chatbots do you prefer to use for university duties or learning purposes? (You may select multiple options.)

□ ChatGPT (1)

□ Google Bard (Gemini) (2)

□ Microsoft Copilot (3)

□ Meta AI (4)

□ DeepSeek (5)

□ Other (please specify—enter all that apply, separated by commas): (6)

____

Display this question:

If Have you ever used a chatbot? (Examples: ChatGPT, Microsoft Copilot, Duolingo Bot, customer servi... = Yes

Prior Experience7 How skilled do you consider yourself in using chatbots for university duties/learning?

○ Not skilled at all (1)

○ Slightly skilled (2)

○ Moderately skilled (3)

○ Skilled (4)

○ Very skilled (5)

Prior Experience8 How confident are you in using chatbots for university duties/learning?

○ Not confident (1)

○ Slightly confident (2)

○ Moderately confident (3)

○ Confident (4)

○ Very confident (5)

UserGuide **To help you get started,** we've prepared a video playlist with step-by-step instructions. This will guide you through everything from creating a chat for your course and chatting with the BuddyGPT. **Please Click to Watch BuddyGPT User Manual Videos When you finish watching the videos, please return the survey. We will give you three different tasks, and you will chat with BuddyGPT about those tasks.**

Q62. Since the system works within the University of Twente's Canvas system, Since the system works within the University of Twente's Canvas system, you must include the you must include the course name (System Design) course name (System Design) in your question to get the correct information. in your question to get the correct information. For example, if you need to find the due date of an assignment, you should ask: "When is the first assignment due?" (Too general, may not work)   "When is the first assignment due for System Design?" (Correct, ensures accurate results)   Once you have completed all tasks, you will be asked to answer questions about your experience with Once you have completed all tasks, you will be asked to answer questions about your experience with BuddyGPT.

End of Block: Interaction Stage Intro

Start of Block: Task1

TimerTask1 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Task1 **Scenario:** Tool Selection You are working on your **System Design project**, which involves creating UML diagrams for a Parking Management System. You are unsure which tool you should use to draw the UML diagrams or how to start with it. Use BuddyGPT to find the tool you can use for your project. How can you use that tool (download/use online)? (Note: UML diagrams are simple visual tools that help organize and plan a system before it is built. The course materials contain information on UML diagrams.) **Feel free to ask BuddyGPT as many questions as you need to complete the task. If you believe BuddyGPT has given a related response that helps you complete the task, you can move on to the next question. You can always return to this scenario later if needed.**

Page Break

TimingTask1Answer Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display this question:

If Scenario: Tool Selection You are working on your System Design project, which involves creating U... Displayed

Task 1 Control. According to BuddyGPT’s response, which tool should you use for your System Design project, and how can you access it? ***(If you are not sure about the BuddyGPT response, please ask again or ask for more details?)***

○ Draw.io, it can be used directly online without downloading and is recommended for project work. (1)

○ Canva, download it using the course-provided software license and install it before use. (2)

○ Visiual paradigm, it should be downloaded and installed using the activation key from the course materials. (3)

○ Eclipse, access it via the course’s project page and install it for online use. (4)

○ I don't know (5)

Page Break

Display this question:

If Scenario: Tool Selection You are working on your System Design project, which involves creating U... Displayed

DifficultyTask1 Please select the option that best describes how difficult did you find this task.

○ Extremely easy (1)

○ Moderately easy (2)

○ Slightly easy (3)

○ Neither easy nor difficult (4)

○ Slightly difficult (5)

○ Moderately difficult (6)

○ Extremely difficult (7)

Task2 **Scenario:** Late Submission You missed an assignment deadline in the **Programming Course** and want to know if late submissions are allowed and whether there are penalties. Use BuddyGPT to find out what to do in this type of situation. Is late submission allowed? How should you submit it? Will there be any penalties? **Feel free to ask BuddyGPT as many questions as you need to complete the task. If you believe BuddyGPT has given a related response that helps you complete the task, you can move on to the next question. You can always return to this scenario later if needed.**

Display this question:

If Scenario: Late Submission You missed an assignment deadline in the Programming Course and want to... Displayed

Task2Control Based on BuddyGPT’s response, what is the recommended action if you miss an assignment deadline in **Programming**? ***(If you are not sure about the BuddyGPT response, please ask again or ask for more details?)***

○ Late submissions are allowed. They have to be submitted on Codegrade as a ZIP file (1)

○ Late submissions are allowed. They have to be submitted as a normal assignment on Canvas (2)

○ Late submissions are not allowed on Canvas. The assignments have to be emailed to the teaching assistants (3)

○ Late submissions are not allowed. The assignments have to be signed off manually by teaching assistants during the lab (4)

○ I don't know (5)

Display this question:

If Scenario: Late Submission You missed an assignment deadline in the Programming Course and want to... Displayed

DifficultyTask2 Please select the option that best describes how difficult did you find this task.

○ Extremely easy (1)

○ Moderately easy (2)

○ Slightly easy (3)

○ Neither easy nor difficult (4)

○ Slightly difficult (5)

○ Moderately difficult (6)

○ Extremely difficult (7)

Task3 **Scenario:** Preparing Project You have a project for your **System Design** course. You have to design a Parking Management System using UML diagrams. You will start by drawing an Activity Diagram. Use BuddyGPT to learn how to start designing the activity diagram for the Parking Management System project. (Note: UML diagrams are simple visual tools that help organize and plan a system before it is built. The course materials contain information on UML diagrams.) **Feel free to ask BuddyGPT as many questions as you need to complete the task. If you believe BuddyGPT has given a related response that helps you complete the task, you can move on to the next question. You can always return to this scenario later if needed.**

Display this question:

If Scenario: Preparing Project You have a project for your System Design course. You have to design... Displayed

Task3Control According to BuddyGPT's response, which of the following statements is right? ***(If you are not sure about the BuddyGPT response, please ask again or ask for more details?)***

○ To start your parking management system, define the key activities based on key processes that may include issuing parking tickets, processing payments etc. (1)

○ To start your parking management system, focus on designing the database first. (2)

○ To start your parking management system, use tools like Eclipse. (3)

○ To start your parking management system, list system users before defining its key processes. (4)

○ I don't know (5)

Display this question:

If Scenario: Preparing Project You have a project for your System Design course. You have to design... Displayed

DifficultyTask3 Please select the option that best describes how difficult did you find this task.

○ Extremely easy (1)

○ Moderately easy (2)

○ Slightly easy (3)

○ Neither easy nor difficult (4)

○ Slightly difficult (5)

○ Moderately difficult (6)

○ Extremely difficult (7)

Post-Test Survey Thank you for participating in this study! In this survey, you will evaluate **BuddyGPT (referred to as "the chatbot" in the questions) based on tasks you completed** Please indicate how much you agree or disagree with the following statements based on your experience with the chatbot. There are no right or wrong answers—please answer honestly based on your experience. ****Once you finish, submit your responses.****

ManualHelpfulness How helpful did you find the user guide in understanding how to interact with the system?

○ Not helpful at all (1)

○ Slightly helpful (2)

○ Moderately helpful (3)

○ Very helpful (4)

○ Extremely helpful (5)

BUS-11 **Based on your experience with the chatbot, please indicate your level of agreement with the following statements.**

	Strongly disagree (1)	Somewhat disagree (2)	Neither agree nor disagree (3)	Somewhat agree (4)	Strongly agree (5)
The chatbot function was easily detectable (e.g., the possibility to modify the settings of the chatbot, make the avatar visible or not, etc.). (1)	○	○	○	○	○
It was easy to find the chatbot. (2)	○	○	○	○	○
Communicating with the chatbot was clear. (3)	○	○	○	○	○
The chatbot was able to keep track of context. (4)	○	○	○	○	○
The chatbot’s responses were easy to understand. (5)	○	○	○	○	○
I found that the chatbot understands what I want and helps me achieve my goal. (6)	○	○	○	○	○
The chatbot gave me the appropriate amount of information. (7)	○	○	○	○	○
The chatbot only gave me the information I need. (8)	○	○	○	○	○
I felt like the chatbot’s responses were accurate. (9)	○	○	○	○	○

I believe the chatbot informs me of any possible privacy issues. (10)	○	○	○	○	○
My waiting time for a response from the chatbot was short. (11)	○	○	○	○	○

End of Block: BUS-11

Start of Block: Post-Ineraction Stage

NPS How likely are you to recommend the chatbot to a friend or colleague?

○ 0 (0)

○ 1 (1)

○ 2 (2)

○ 3 (3)

○ 4 (4)

○ 5 (5)

○ 6 (6)

○ 7 (7)

○ 8 (8)

○ 9 (9)

○ 10 (10)

Appendix C: Scales of BUS-11's Four Factors

Scales of BUS-11	Very poor	Poor	Average	Good	Very good
1. Accessibility	0 to 3.5	> 3.5 to 3.7	> 3.7 to 3.92	> 3.92 to 4.19	> 4.19 to 5
2. Functional Inter. conv	0 to 3.32	> 3.32 to 3.53	> 3.53 to 3.68	> 3.53 to 4.05	> 4.05 to 5
3. Privacy	0 to 2.52	> 2.52 to 2.63	> 2.63 to 2.77	> 2.77 to 3.18	> 3.18 to 5
4. Responsiveness	0 to 4.03	> 4.03 to 4.23	> 4.23 to 4.38	> 4.38 to 4.58	> 4.58 to 5
Scores (%)	0 to < 67%	≥ 67 to < 71%	≥ 71 to < 74%	≥ 74 to < 80.1%	≥ 80.1%

Appendix D: The Items of the Four-Factors BUS-11

Factors	Items
1. Accessibility — Perceived accessibility to chatbot functions	1. The chatbot function was easily detectable. 2. It was easy to find the chatbot
2. Functional Interactive Conversation — Perceived quality of chatbot functions, conversation, and information provided	3. Communicating with the chatbot was clear
	4. The chatbot was able to keep track of context 5. The chatbot's responses were easy to understand 6. I find that the chatbot understands what I want and helps me achieve my goal 7. The chatbot gives me the appropriate amount of information 8. The chatbot only gives me the information I need 9. I feel like the chatbot's responses were accurate
3. Privacy — Perceived privacy and security	10. I believe the chatbot informs me of any possible privacy issues
4. Responsiveness — Time response	11. My waiting time for a response from the chatbot was short

Appendix E: The content and explanation of Tasks

- Task 1 (Tool Selection): The participant, acting as a student working on a project, needed to design Unified Modeling Language (UML) diagrams for a Parking Management System. The task required the participant to query BuddyGPT to identify which tool could be used to draw the UML diagram.
- Task 2 (Late Submission): The participant, as a student who had missed an assignment, had to determine whether late submissions were allowed and if any penalties applied. BuddyGPT was expected to address this inquiry.
- Task 3 (Preparing Project): The participant needed to learn how to start and draw an activity diagram (one of the UML diagrams) for the Parking Management System project by interacting with BuddyGPT.

Appendix F: The table of six tasks order group sequences

Order	Task 1	Task 2	Task 3
1	1	2	3
2	3	2	1
3	2	3	1
4	2	1	3
5	3	1	2
6	1	3	2