Feasibility and Acceptance of Chatbots Embedded in Healthcare Curricula:

Matthew Pears, Eirini Schiza, James Henderson, Natalia Stathakarou, Klas Karlgren, Panagiotis. D. Bamidis, Iraklis Tsoupouroglou, Constantinos. S. Pattichis, and Stathis. Th. Konstantinidis

CEPEH Report

December

2022

see CEPEH.eu for more information

Acknowledgements

This work is supported by the ERASMUS+ Strategic Partnership in Higher Education "Chatbot Enhance Personalise European Healthcare Curricula (CEPEH)" (www.cepeh.eu) (2019-1-UK01- KA203-062091) project of the European Union.

Thank you to Mohammed Roshan for assistance in R data processing.

The CEPEH Team

Abstract

This document details the evaluation of each resource in terms of the feasibility and acceptance from the end-users. There was evidence of identifying the feasibility of such resources into formal training and studies exist on the acceptance of such resources, with promising results. However, all these studies defined the need for further research in the area until the use of chatbots in healthcare education became common. Furthermore, the creation process of CEPEH resources was significantly different and had improvements to current methods, due to the co-creation process, and use of low cost but effective technology.

Contents

List of Figures

List of Tables

List of Abbreviations

CEPEH . . . . Chatbot Enhance Personalised European Healthcare curricula

HELM . . . . . Health E-Learning and Media team

RLO . . . . . Reusable Learning Object

ASPIRE . . . Aims Storyboarding Production Implementation Release Evaluation

NLP . . . . . Natural Language Processing

NLU . . . . . Natural Language Understanding

A.I . . . . . Artificial Intelligence

TAM . . . . . Technology Acceptance Model

SUS . . . . . System Usability Scale

CUQ . . . . . Chatbot Usability Questionnaire

HIG . . . . . HELM is Great

Introduction

Personalised Healthcare Education is needed to meet growing demand and quality maintenance. There is a growing evidence around chatbots, namely machine conversation systems- these programs have the potential to change the way students learn and search for information.

Chatbots can quiz existing knowledge, enable higher student engagement with a learning task, or support higher-order cognitive activities. In large-scale learning scenarios with a hight student-to-lecturer ratio, chatbots can help tackle the issue of individualized student support and facilitate personalised learning. However, limited examples of chatbots in European Healthcare Curricula have been utilised to combine both the continuum of cognitive processes presented in Bloom's taxonomy, with the idea that some repetitive tasks can be done with a chatbot- to provide greater access or to scale faculty time.

Thus, CEPEH strategic partnership has co-created open access chatbots utilising artificial intelligence, promoting innovative practices in digital era, by supporting current curricula and fostering open education.

CEPEH Erasmus+ strategic partnership aimed to co-design and implement new pedagogical approaches and, in particular, chatbots for European medical and nursing schools. CEPEH used use participatory design to engage stakeholders (students, healthcare workforce staff, lecturers, clinicians, etc.) in order to co-design effective chatbots and release them as open access resources. Through CEPEH, effective use of digital technologies and open education were be incorporated into healthcare curricula. This enabled students to increase their health and medical related skills through flexible learning.

CEPEH expected that students adopted this new digital pedagogy and improve their skills and competences through flexible personalised learning, while the teaching staff enhanced their e-learning tool co-creation competences and make use of co-design best practices and recommendations for use. It is also expected increased cooperation between the partners. Thus, in the long term, CEPEH expects to influence the development of medical and nursing curricula with this digital innovation, foster the quality of the future healthcare workforce and further improve international competitiveness of the partners' healthcare curricula. This document details the evaluation of the resources created by the CEPEH team.

The evaluation specifically explored the feasibility and acceptance from the end-users. These end-users are learners in European healthcare higher education institutions.

There was firstly evidence for the need to identify the feasibility of chatbots and similar resources into formal education and training, with a further need to improve access to these types of learning resources. Of course, studies exist on the acceptance of chatbots, virtual patients, and many other healthcare applications, with promising results. However, through various limitations, we believed there was further research to be completed to accelerate the design, development, implementation, and evaluation processes. These have financial, stakeholder, time, and efficacy benefits. The creation process of CEPEH resources was significantly different to most in the literature, and this report highlights the approach of the CEPEH team towards enhancing personalised healthcare education can be achieved.

Background

The working practices of CEPEH are aimed at maximizing efficacy of these chatbots as learning resources, and provided a sense of shared development and ownership from all stakeholders. The process normally begins with workshops in which the project is scoped and team building occurs. The CEPEH workshops involve the widest possible team of stakeholders including tutors, students, healthcare

workers, learning technologists, health service users and carers- depending on the materials being created.

For readers who are interested in using these high quality digital resources please access them for free at CEPEH.EU

The next section will now present the evaluation of all CEPEH chatbot resources.

1 Method

1.1 Participants

This dataset had 14 males and 28 females therefore a total of 42 participants. It was a repeated measure design whereby each participant used the 4 chatbots developed by the CEPEH team. Therefore, there are 42 points of data in the condition before testing, and 126 data points after testing the chatbots- for a total of 168 row of data. There were 78 questions asked in total, therefore the full dataset had approximately 6000 cells recorded.

There were 22 females and 7 males from Greece. There were 3 females and 4 males from Cyprus. There were 2 females and 2 males from Sweden, and there were 2 participants from the United Kingdom (see(**??**).

The majority 36 participants, were student, with 3 being learning technologists, 2 were lecturers, and 1 was a doctor. Although there could be a difference in these groups, the design was within- groups therefore each participants pre-usage metrics were the comparative control data, and participant differences did not affect the evaluation.

Figure 1.1: Location and Profession of Participants

1.2 Procedure

For each resource created by the Partners, the same experimental methodology was followed. For each resource created by partners, students performed a study within an online or face to face workshop or course. Student participants joined from Greece, Cyprus, Sweden, and the United Kingdom. A repeated measures design was used as the same group measures were taken before and after usage of the chatbots. They were recruited via staff members in the CEPEH group.

Participants were asked prior to the study if they agree to participate, providing them with a PIS form. Participants had the opportunity to discuss with the research team prior to the study and before consent is given. Then, participants used the chatbot resources independently and technical support was provided. Finally, post-intervention measures were recorded.

Some of the participants were invited to participate in Focus Group Discussions (FGD), and each FGD lasted between 15 to 25 minutes, with 5-10 participants. Participants were asked if they would like to be informed of the findings of the study.

1.3 Design

The data captured from the participants were their initials and numerical day of birth, used as anonymous identifier for pre-post analysis. Their institution was captured (Aristotle University of Thessaloniki, CYENS Centre of Excellent, Karolinska Institute, and The University of Nottingham), and Sex (Male/Female/Other).

Before any interaction with the learning resources, various perceptions of chatbot such as confidence and easy of use, usefulness, Influence from others, and current learning resources (videos, textbooks, Google, friends etc), were captured. Descriptive data was produced alongside repeated measures t-tests. Repeated measures t-tests were the appropriate test to use as this explores differences between groups, there were no covariates and we did not have several dependant variables. There was one Independent factor being Chatbot use having 2 levels (pre/post). There were 3 chatbots therefore there was option for ANOVA to determine where differences lie if statistical differences were found however this was not wholly appropriate for the data type and not necessary for pre-post comparison.

1.4 Materials and Measures

The measures used fit within a newly developed Chatbot Evaluation Framework-which takes the best measures of 5 previous frameworks. Denecke and Warren [2] derived several quality dimensions and attributes from previous chatbot literature. They formed six perspectives from their review of articles and mobile health applications.

These six perspectives were: 1) Task-oriented, 2) Artificial intelligence, 3) System quality perspective, 4) Linguistic perspective, 5) UX Perspective, 6) Healthcare quality perspective.

To capture these perspectives, we used several validated materials that can distinguish these elements of the CEPEH chatbots.

1.4.1 Chatbot Usability Questionnaire (CUQ)

The Chatbot Usability Questionnaire (CUQ) [4] is a new questionnaire specifically designed for measuring the usability of chatbots by an interdisciplinary team from the Ulster University. CUQ can be used alongside the prevalent System Usability Scale Score (SUS) [5]. Multiple metrics are more appropriate when measuring usability of chatbots [6] therefore a combination of two scores can provide an all-inclusive overview.

1.4.2 UTAUT2 (Unified Theory of Acceptance and Use of Technology)

The underpinning theory of the UTAUT2 is that there are four key constructs to the intentions of using technology based resources: 1) performance expectancy, 2) effort expectancy, 3) social influence, and 4) enabling conditions.

The TAM and the UTAUT2 have cross over in measuring technology acceptance, however the UTAUT2 has more applied probing questions. Few studies exist that use technology acceptance theories for the intention to use products that explicitly incorporate AI. A recent extension of the UTAUT2 model added five (health, convenience comfort, sustainability, safety, security, and personal innovativeness) additional influencing factors to accommodate for AI [7]. This can be used for products in either health, household use, or mobility and can help to explain behavioural intention and use behaviour of chatbots.

1.4.3 System Usability Scale

The System Usability Scale (SUS) was used [10] and is a widely used and adopted usability questionnaire. It is popular due to its unbiased and agnostic properties, a non proprietary, and a quick scale of 10 questions. However, as there are the CUQ and parts of the UTAUT2 we have selected only 2 questions which do not

cross-over with the other measures. These are improtant statements however and good indicators of usability when assessed with the other results.

1.4.4 Computer Self-Efficacy Scale Tool

The 10 question CSEST is based on the 32-item questionnaire by Murphy, Coover, and Owen (1989). It can be adapted for any technology and we have selected only a few pertinent questions. Participants are asked to think about using the CEPEH chatbots and answer that they would use the chatbots if I had never used a product like it before; If they could call someone for help if I got stuck, or if someone showed them how to do it first, and other similar usage questions.

1.4.5 Technology Acceptance Model (TAM)

The Technology Acceptance Model (TAM) [1] was specifically developed with the primary aim of identifying the determinants involved in computer acceptance in general; secondly, to examine a variety of information technology usage behaviours; and thirdly, to provide a parsimonious theoretical explanatory model.

TAM suggests that attitude would be a direct predictor of the intention to use technology, which in turn would predict the actual usage of the technology. The only modification to the nine sub-scales of the questionnaire consists of applying the items to the context of chatbots. All the items, except those measuring attitudes, utilize a seven-point Likert scale ranging from “strongly agree” to “strongly disagree” with a middle neutral point [2].

The nine sub-scales of the questionnaire:

Ease of use of chatbots Perceived usefulness of chatbots Intention of use. Attitude toward usage of chatbots. Perception of personal efficacy to use a chatbot resource. Perception of external control toward chatbots. Anxiety toward chatbot use. Intrinsic motivation to use chatbot resources. Perceived costs of chatbots.

1.4.6 Qualitative Measure- Focus Group Discussions

Focus groups are a pervasive means of market research and provides credible acceptance evaluators regarding the penetration that a product or service will have on a target demographic.Focus groups are a form of qualitative research consisting of interviews or structured discussions, in which a group of people are asked about their perceptions, opinions, beliefs, and attitudes towards a product, service, concept, advertisement, idea, or packaging.

Questions are asked in an interactive group setting where participants are free to talk with other group members.During this process, the researcher either takes notes or records the vital points he or she is getting from the group. Researchers select members of the focus group carefully for effective and authoritative responses.Relevant stakeholders, then, can use the information collected through focus groups to receive insights on a specific product, issue, or topic focus [7].

A series of short focus group sessions identified the feasibility of CEPEH resources for formal curricular integration. These sessions, spanning no more than 1-1.5 hours and consisting of no more than 5-7 persons each explored all axes of curricular integration such as accessibility in the classroom, use case scenarios, technology requirements for curricular integration etc.These axes were formalized by the research team, in each evaluation site, to consider the curricular details of each institution.

Figure 1: Flow diagram of the recruitment process

2 Results

2.1 Participants' Characteristics

When participants were asked the amount of time they have used a chatbot in any form or subject, 23 stated they had never used a chatbot. Further, 19/42 stated having used a chatbot at least once for between 0-4 hours of use in total. These are likely commercial/website- based assistant chatbots however there are some medical/healthcare resources known to be used in anatomy and/or patient interactions. One individual had spent much longer time with usage- this was the mature student.

Table 2.1: Previous Chatbot Usage of Participants

Previous_Chatbot_Usage	n
1-4 hours	15
10-19 hours	1
20+ hours	1
5-9 hours	2
Never	23

In short, approximately 50% had never used a chatbot, and 45% had used a chatbot, at some period over the years, for a short period of time.

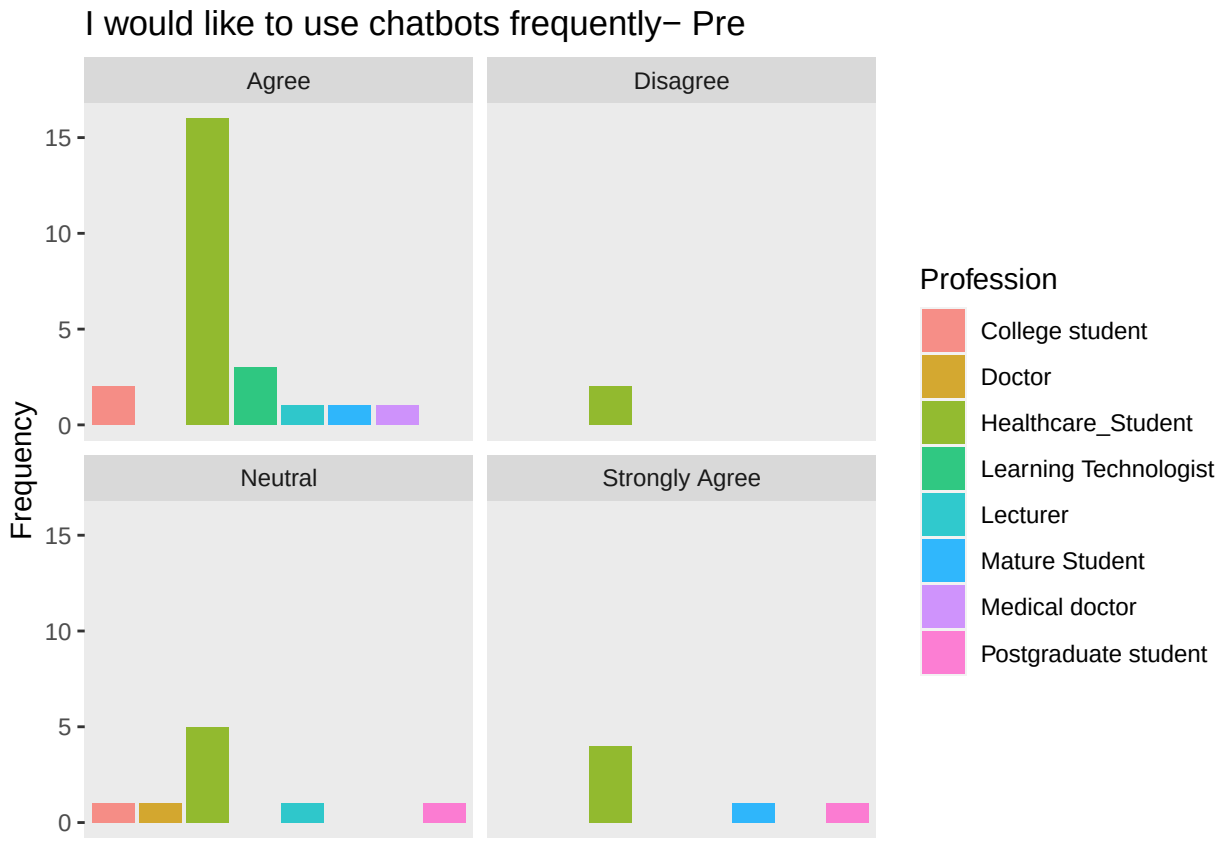


Figure 2.1: Chatbot Usage History- Pre

Most learners use books or online books as resources. They may use multiple sources however they were asked to note the primary source. Only 6 stated their primary sources were *Online videos/interactive materials* which includes such tools as chatbots.

The first boxplot (**??**) shows learners perceptions of easy of use of mobile app and other educational mobile resources

(**??**) shows the opinions of all participants on the usefulness of chatbots. Many had not had experience with them yet had positive rating.

This positive opinions of chatbots may be from colleagues, friends, media, tutors, or other social information of the benefits in healthcare education. Around 25% were neutral or disagreed that healthcare chatbots were useful.

The participants then used the 4 chatbots, and completed the post-usage survey after each chatbot. Results after use are as followed:

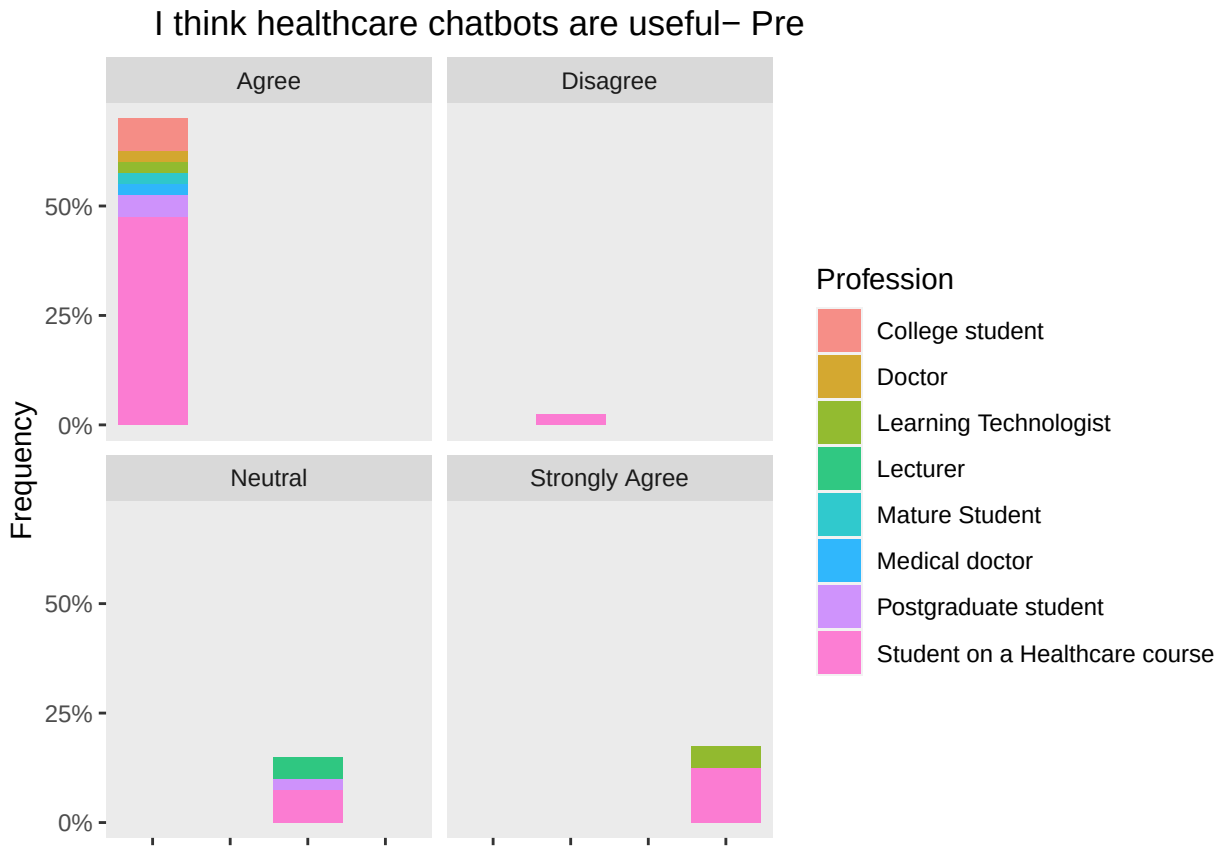


Figure 2.2: Chatbots are Useful Opinion- Pre

2.2 Chatbot Usability Questionnaire (CUQ)

2.2.1 CUQ Calculation tool

The CUQ was developed by researchers at Ulster University, Link and as the calculation can be complex, a dedicated calculation tool has been created.

Please download the CEPEH CUQ calculation tool which has all of the data entered, so you can see the CEPEH CUQ scoring

Click here to download CUQ calc tool

Click here to download CEPEH CUQ score result

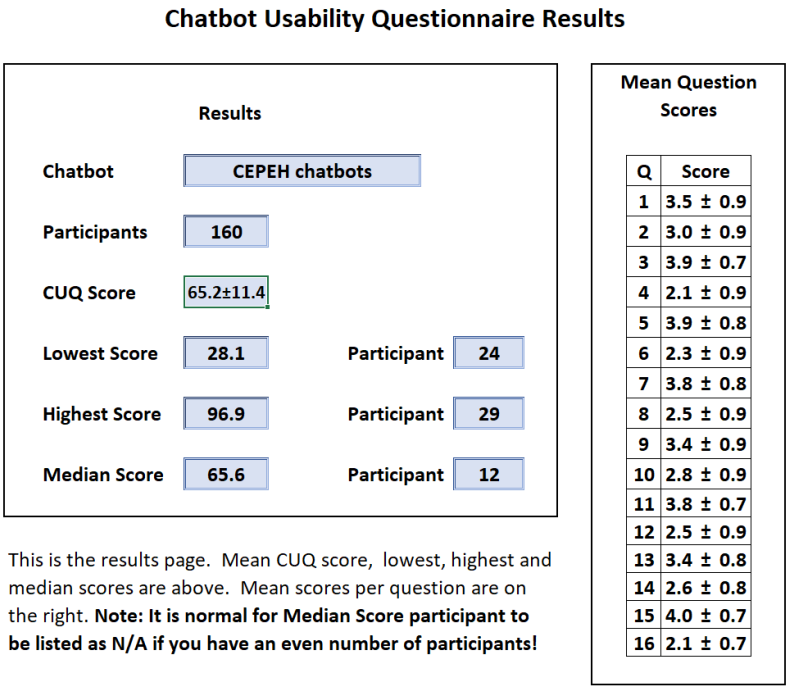


Figure 2.3: CUQ CEPEH Score

Although the design and development was similar, each chatbot CUQ score was calculated to understand how the topic content may affect usability:

The breakdown of the chatbots was:

- Aristotle University of Thessaloniki CUQ score = 63/100
- CYENS Centre of Excellence CUQ score = 67/100
- Karolinska Institute CUQ score = 63/100
- University of Nottingham CUQ score = 68/100

The score for all 3 chatbots grouped was 65/100. See Discussion CUQ section for interpretation

Figure (**??**) shows the CUQ scores as a scatter plot to highlight how there was a moderate distribution of results. Further exploration is required to understand which elements are causing this spread, and if it was due to problems within a small group of learners.

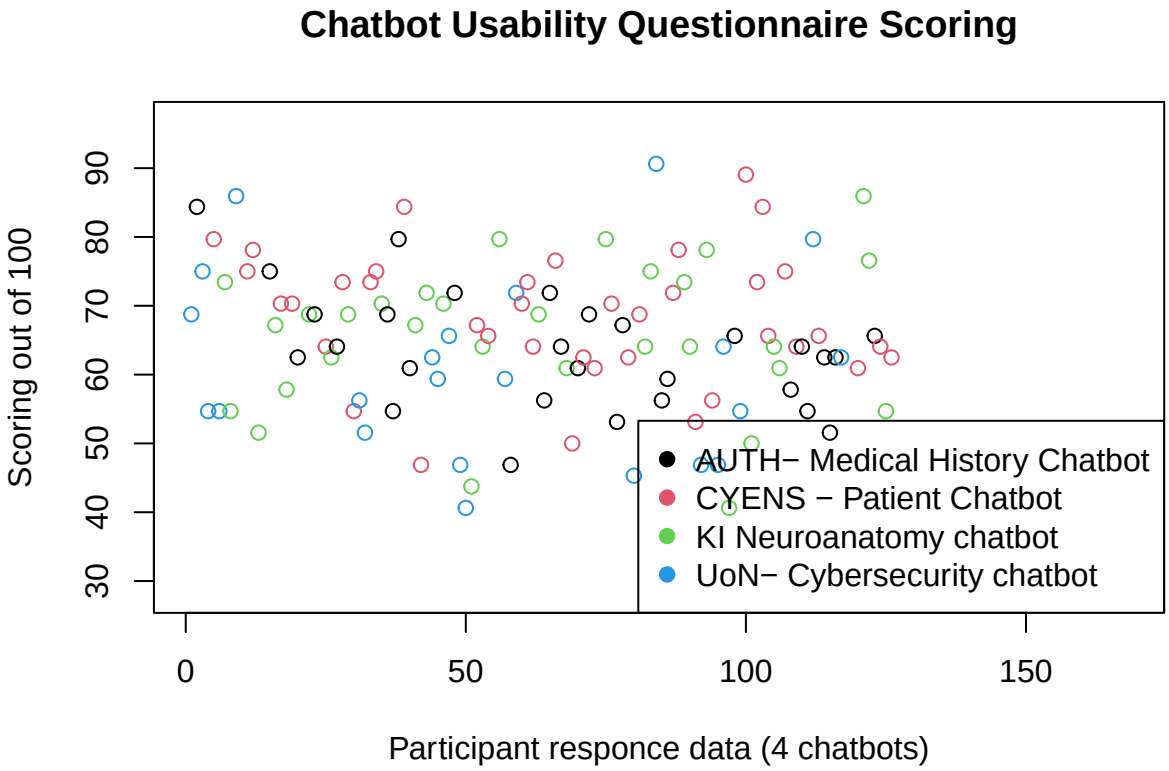


Figure 2.4: CUQ Scatter Plot

2.3 System Usability Scale (SUS) Scores

Note= The amount of 'agreement' is defined as the addition of 'Agree' and 'Strongly agree' responses.

The SUS score should consist of 10 items. However, some SUS questions were improved upon by 1 or more CUQ questions, specifically to this Chatbot study. The SUS results would be obscured by the CUQ scores, expect 2 that did not have cross-over. The two questions were:

- I would like to use the CEPEH chatbot I tested, more frequently (SUS1)(post)
- I felt confident using the CEPEH chatbot (SUS2)(post)

This meant the score of the SUS was not created, however the CUQ score better represented the Learners' perceptions of the CEPEH chatbot in terms of feasibility of use and acceptability in healthcare curricula.

Keep Using CEPEH Chatbot	Responces
Agree	66
Disagree	15
Neutral	17
Not Applicable	3
Strongly Agree	23
Strongly Disagree	2

The table (@ref(tab:SUS keepusing)) above shows the results for agreement participants may continue to use the CEPEH chatbots: 89/126 (70%) agreed or strongly agreed. However, there were 23 records that learners were neutral or disagree they would continue use.

Confidence using CEPEH Chatbot(s)	Responces
Agree	71
Disagree	11
Neutral	21
Not Applicable	4
Strongly Agree	19

Confidence when using the chatbots is in table (**??**)- it shows the distribution of agreement for participants for all 4 chatbots. The table shows 90/126 records that participants feel they are confident in using the chatbots. However, 21/126 (16%) were neutral and 11/126 (8.5%) disagreed and this was explored in the qualitative analysis section.

2.4 Technology Acceptance Model

The TAM questions were analysed according to their subsets. The subsets were Perceived Usefulness (PU) and Perceived Easy of Use (PEU)

The questions were: Perceived Usefulness (PU)

1. Using CEPEH chatbots would enable me to accomplish tasks more quickly
2. Using CEPEH chatbots would increase performance
3. Using CEPEH chatbots would increase my productivity

4. I would find CEPEH chatbots useful on my course

Perceived Easy of Use (PEU)

5. Learning to use CEPEH chatbots would be easy to me
6. It would be easy for me to be skilful at using CEPEH chatbots
7. My interactions with CEPEH chatbots would be clear and understandable
8. I would find CEPEH chatbots easy to use

Results

The scores as a percentage of agreement, were calculated by averaging the subsets and interpreted as:

- Before using the CEPEH chatbots, there was 66% (2.2/5) agreement for the Perceived Usefulness of chatbots in healthcare education, and after 48% (2.6/5) agreed.
- Before using the CEPEH chatbots, there was 64% (2.3) agreement for Perceived Ease of Use of chatbots in healthcare education, and after 51% (2.56) agreed.

The justification for this may be due to being early versions of applications with limited functionality and functions which can be difficult for user to experience the intended further range of features and learning exercises.

2.4.1 Knowledge and Trust after Use

CYENS chatbot had around 10 more participants stating that they were neutral on gaining knowledge of the topic

The figure (@ref(fig:Boxplot trust)) shows the ratings by participants of the CEPEH Chatbots to provide them with the necessary course information.

This is a integral element in learners' motivational and educational choices to reuse the learning resources. As previously described, the trust of the information is also a factor in these responses.

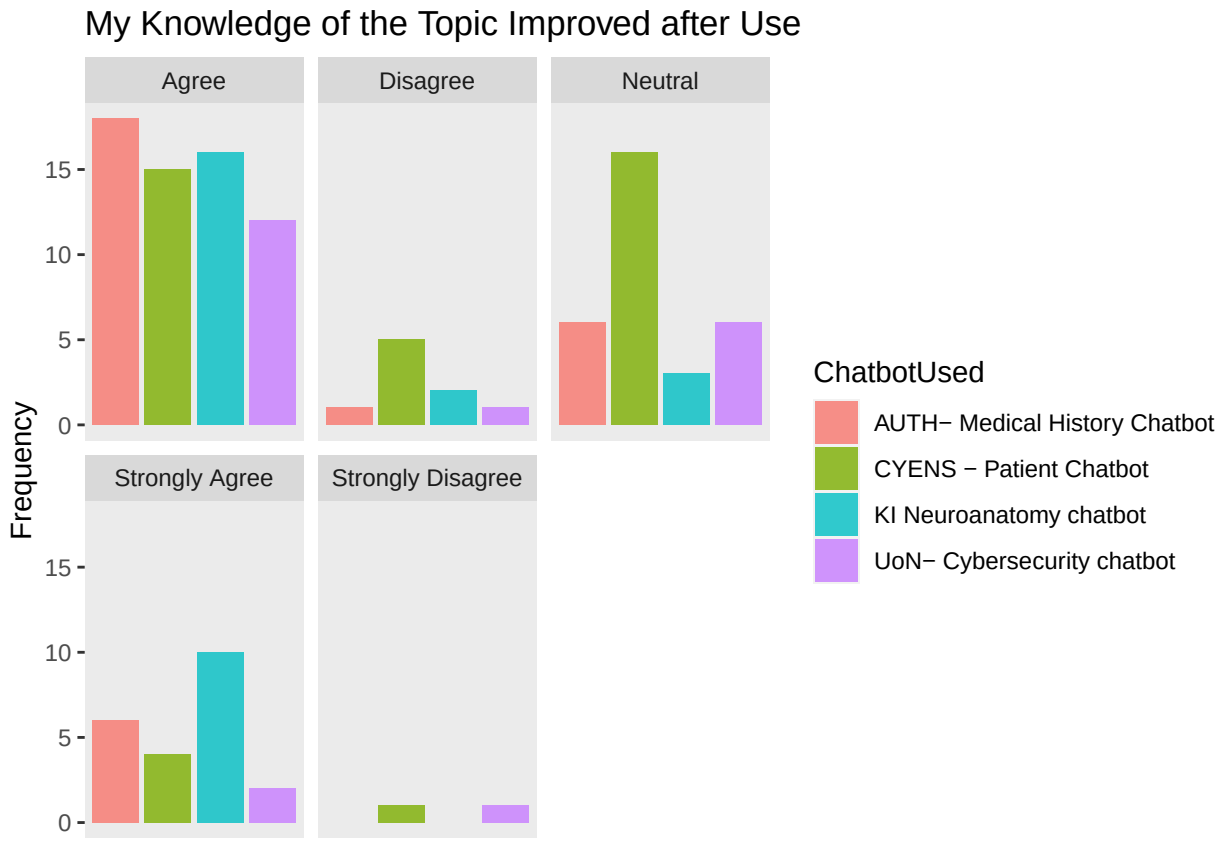


Figure 2.5: Improvements in Knowledge
(#fig:Boxplot knowledge)

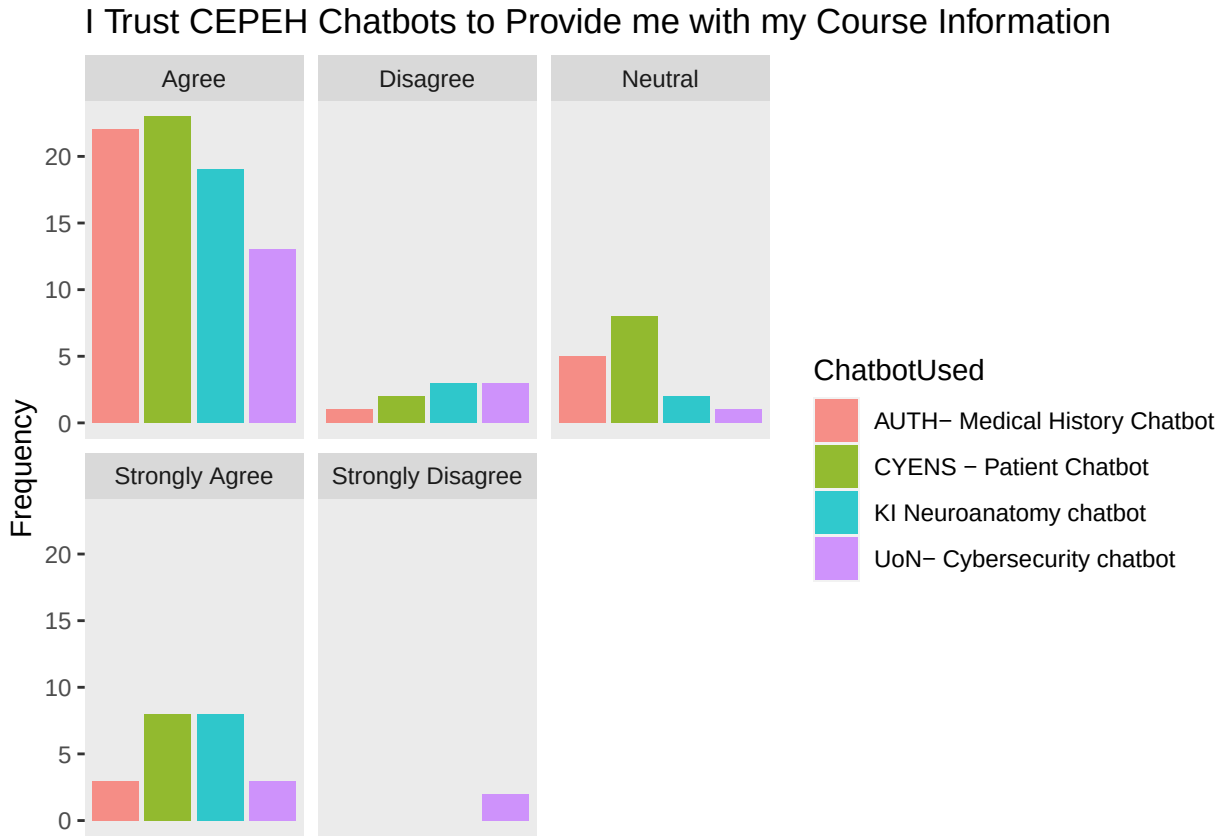
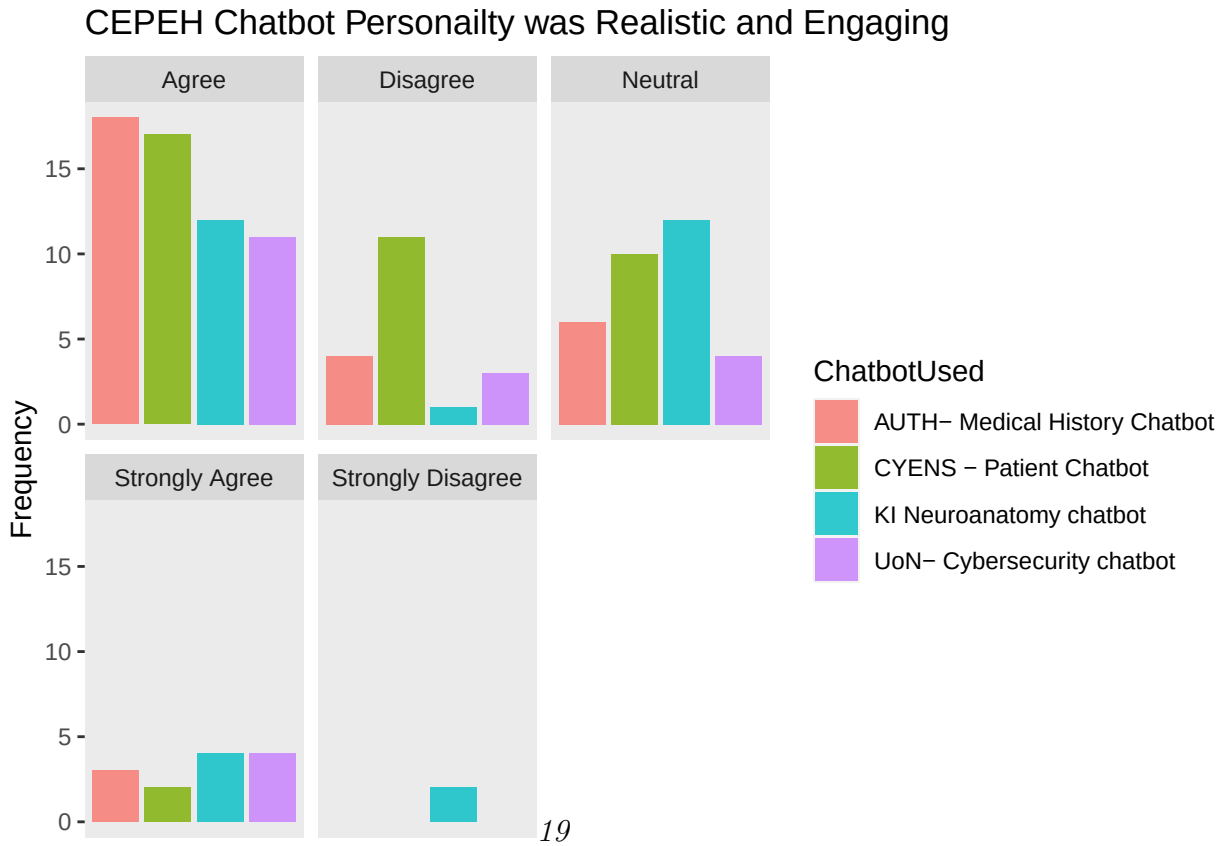


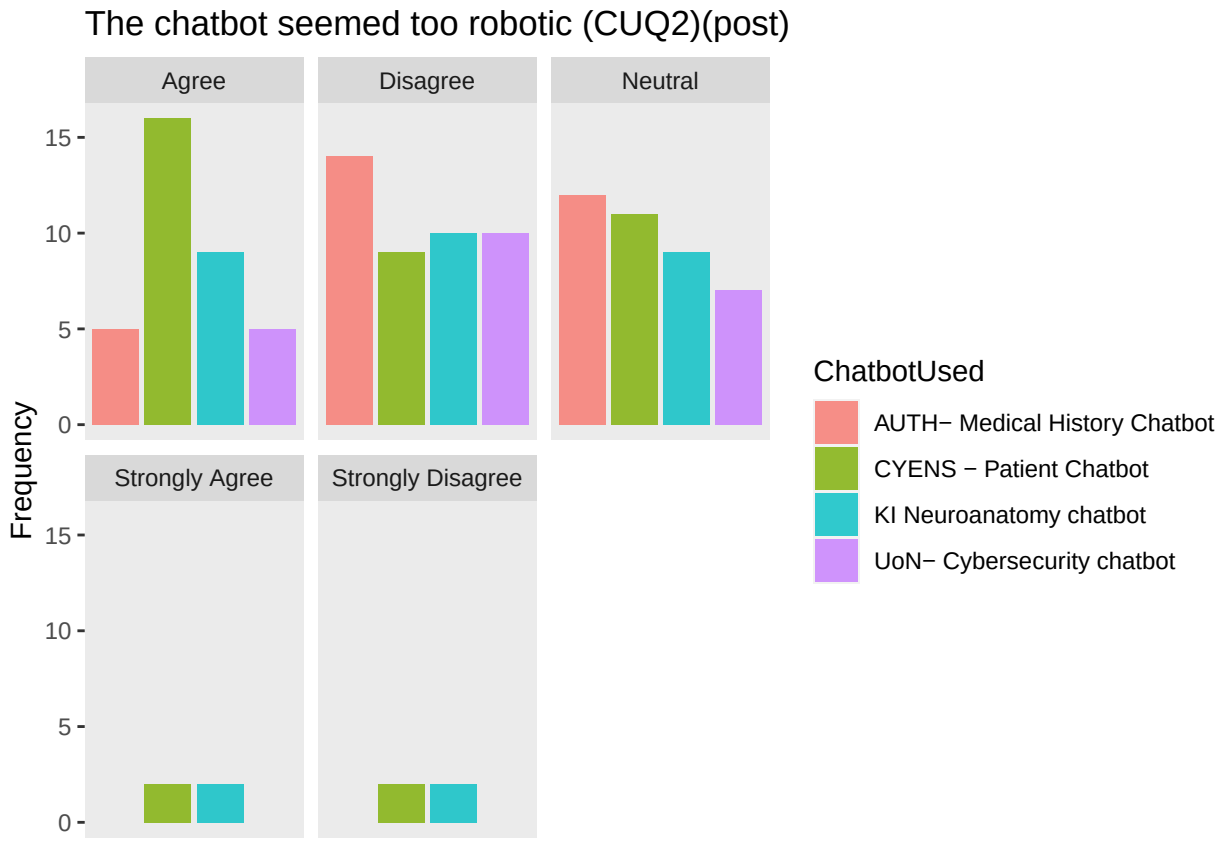
Figure 2.6: Trust Chatbots POST use

(#fig:Boxplot trust)



There was mixed results for the chatbot used being realistic and engaging. This question has two descriptive terms however based on the other results we understand that the chatbots' NLP logic, or ability to respond required improvement to be more 'smooth' in replying. The primary limitation was found in the 'robotic' interactions(See Figure x). This was investigated further in the 'Text Mining' and 'Sentiment Analysis' sections.

2.4.2 Personality and Interactions



The chatbot seemed too robotic results had the largest mix of responses, and for all 4 chatbots evaluated. The University of Nottingham Cybersecurity chatbot had more deterministic pathways with exploitation of the NLP modelling to provide illusion of realism. This may explain why there was less agreement. However, Neutrality and/or agreement was not desired.

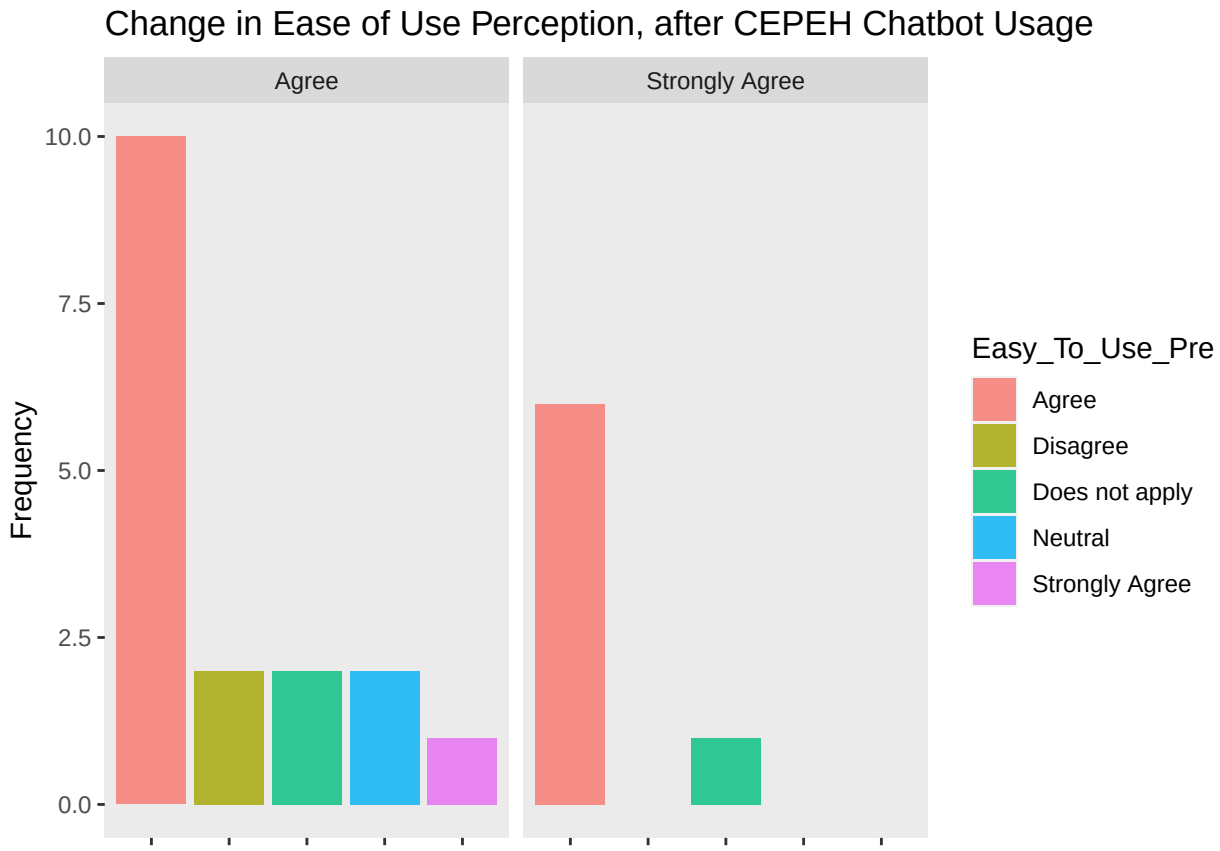


Figure 2.7: Ease of Use Comparison

2.4.3 Ease of Use and Seeking Support

After usage, there was only agreement in Ease of Use- as shown in (**??** as there are no 'Neutral' or disagree columns. Any learners with disagreement before using the CEPEH chatbots, after believed they were easy to use.

Those who disagreed or were neutral in the pre usage measure, improved their understanding that help was available with the CEPEH chatbots. After usage, 40 participants agreed they could get help if they had difficulty using the resources.

2.5 Inferential Statistics

2.5.1 Repeated Measures T-test results

After using the CEPEH chatbots, majority of participants stated they would reuse the chatbots. However, there was 6 counts of *disagree* or *strongly disagree* for