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.

The CEPEH Team

Abstract

Healthcare education can be supported by machine learning conversation agents. However, there is rapid pace of technical development, complex subject areas in healthcare and sensitive design and development protocols which required expertise. With these issues, current outcomes have barriers in design, development, implementation, and in cases their evaluation. By utilizing a long-standing framework named the ASPIRE framework, the CEPEH team have reinvented parts of the process to streamline and dampen common problems. Stakeholder inclusion was facilitated by the ASPIRE process and the synergistic effect of development heuristics, learners' perspectives, and subject expertise validation. The resultant 4 chatbots, in differing healthcare topics, were evaluated to understand how this simplistic and inclusive approach changes learners' perspectives and experience of chatbots/conversational agents, to promote uptake and course performance. The results showed the majority of descriptive metrics (means, medians, modes) improved marginally, with minority showing no or reduced feedback. Majority consensus of improved experience and perspective is a great outcome considering learners' concerns and anxieties noted in previous literature. This heterogeneity was explored to understand the technical and usage limitations in some users. The CEPEH team concluded that the processes, frameworks, development tools, and evaluation metrics can improve and encourage researchers, learning technologists, educators, and students to produce similar supportive, intuitive, accessible, and easily sustainable learning resources.

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 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 high student-to-lecturer ratio, chatbots can help tackle the issue of individualized student support and facilitate personalized learning. However, limited examples of chatbots in European Healthcare Curricula have been utilized 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 utilizing articial intelligence, promoting innovative practices in the 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 participatory design to engage stakeholders (students, healthcare workforce sta, lecturers, clinicians, etc.) in order to co-design eective chatbots and release them as open access resources. Through CEPEH, effective use of digital technologies and open education were incorporated into healthcare curricula. This enabled students to increase their health and medical related skills through exible learning.

CEPEH expected that students adopted this new digital pedagogy and improve their skills and competences through exible personalized learning, while the teaching sta enhanced e-learning co-creation competences and make use of co-design best practices and recommendations for use. It was also expected that increased cooperation between the partners would occur. Thus, in the long term, CEPEH expects to inuence 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 specically explored the feasibility and acceptance from the end-users. These end-users are learners in European healthcare higher education institutions.

There was rstly 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 nancial, stakeholder, time, and ecacy benets. The creation process of CEPEH resources was signicantly dierent to most in the literature, and this report highlights the approach of the CEPEH team towards enhancing personalized healthcare education.

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.

oh and by the way, this site updates with every refresh and processes new information. Try it yourself and refresh this page to recieve a new compliment everytime: **You are primo!**

1 Method

Contents

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.

Figure 1.1: Location and Profession of Participants

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(1.1).

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.

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. Student participants joined from Greece,

1. Method

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 sta 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 ndings of the study.

Figure 1: Flow diagram of the recruitment process

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 ease 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 and Wilcoxon signed rank test. Repeated measures paired sample 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, with 2 levels (pre/post).

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 questions from multiple measures can help internal consistency and interpret the results more approporiately.

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 important 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 data collection for research and provides end user insights regarding the penetration, usage, scope, and impact of a learning resource. 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 the item of interest. 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 use the information collected through focus groups to receive insights for improvements [7].

A series of short focus group sessions identified the feasibility of CEPEH resources for formal curricular integration. These sessions, spanning no more than 30 minutes each and consisting of no more than 5-7 persons, 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.

2
Results

Contents

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.

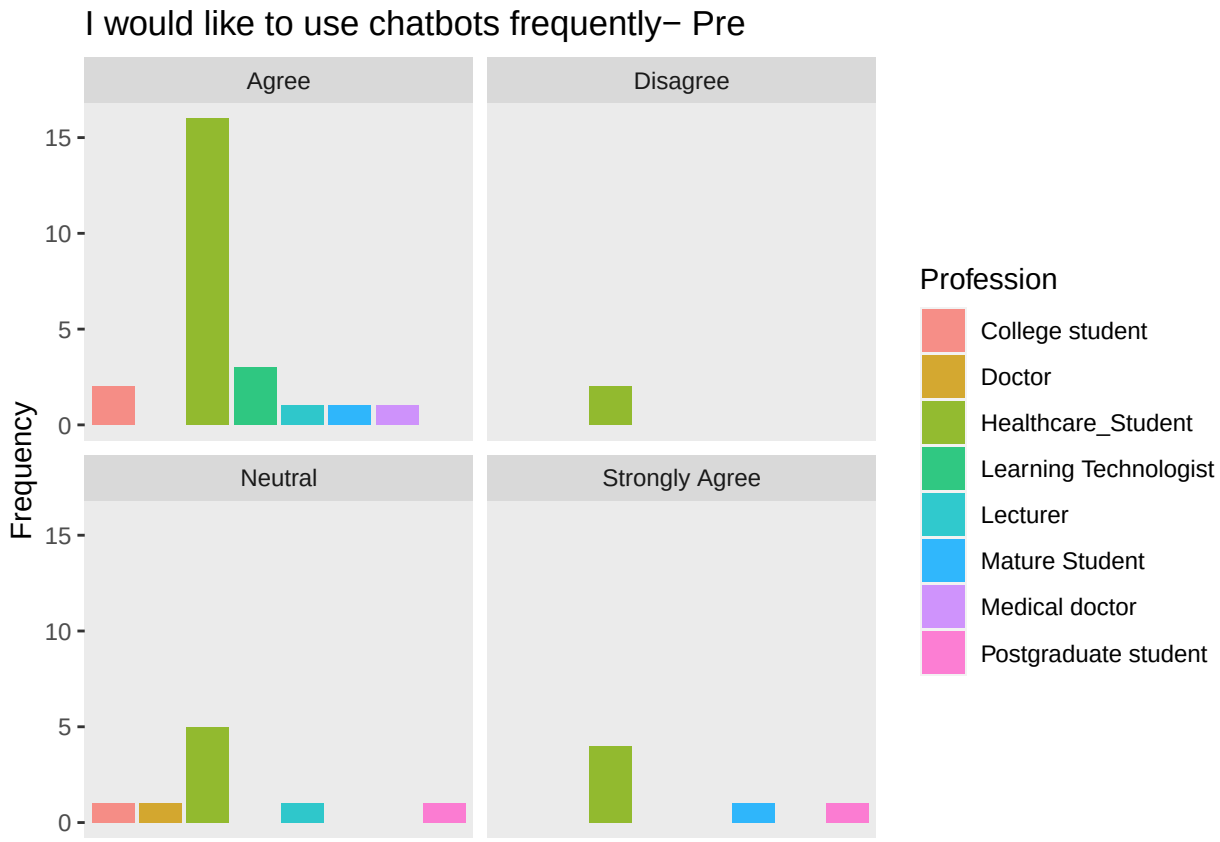


Figure 2.1: Chatbot Usage agreements- Pre

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.

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 (2.1) shows the intention, or at least the learners interest in

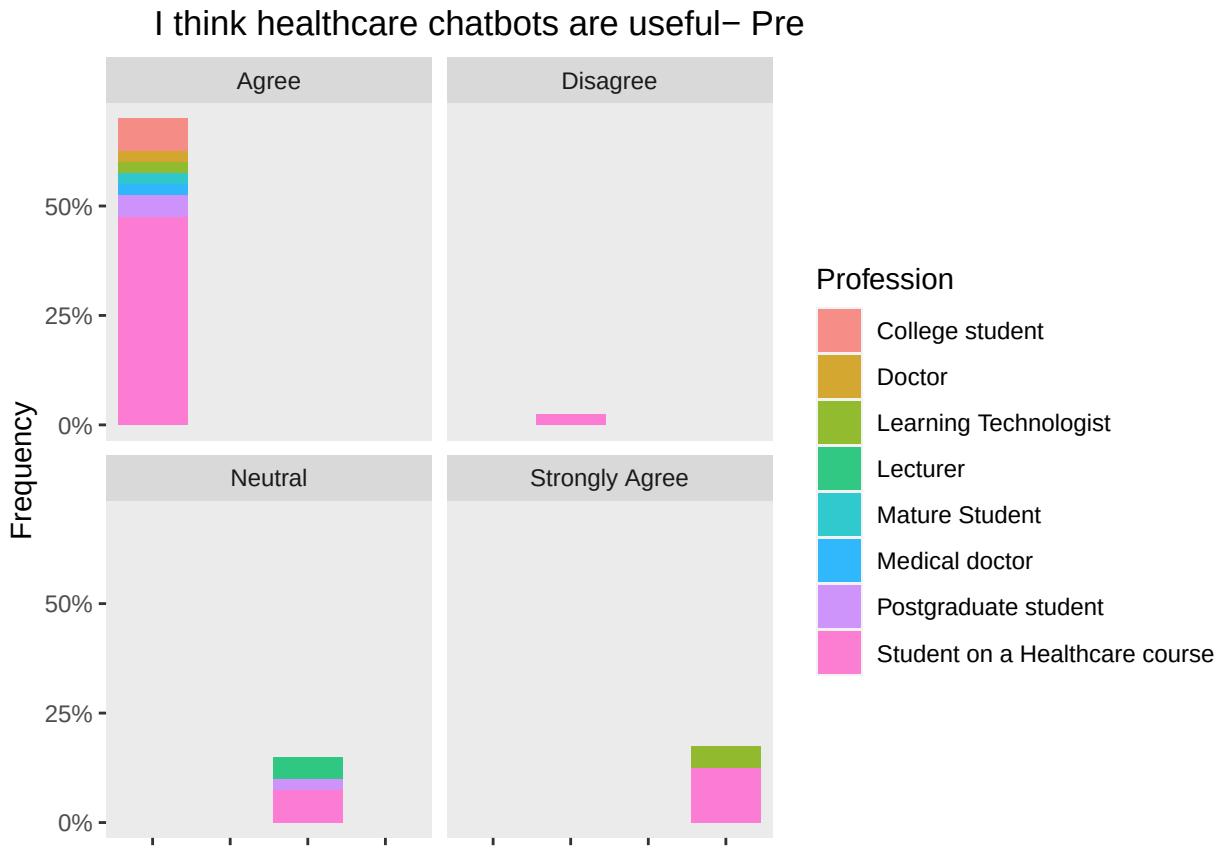


Figure 2.2: Chatbots are Useful Opinion- Pre

using chatbots by means of agreeing that they would like to use chatbots if they had the opportunity. 20 healthcare students agreed (16) and strongly agreed (4) which made about 50% of participants. 9 were neutral with 1 disagree.

(2.2) 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:

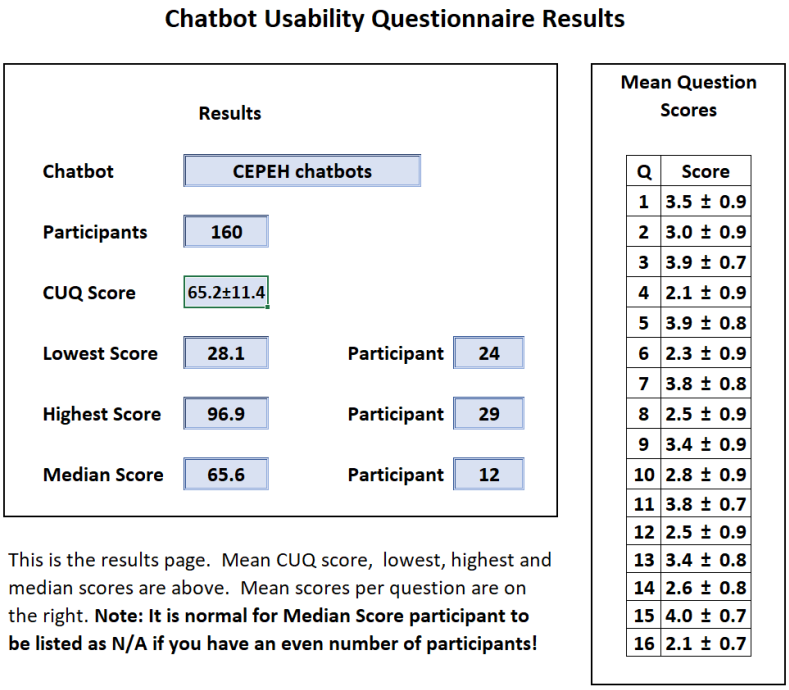


Figure 2.3: CUQ CEPEH Score

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

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

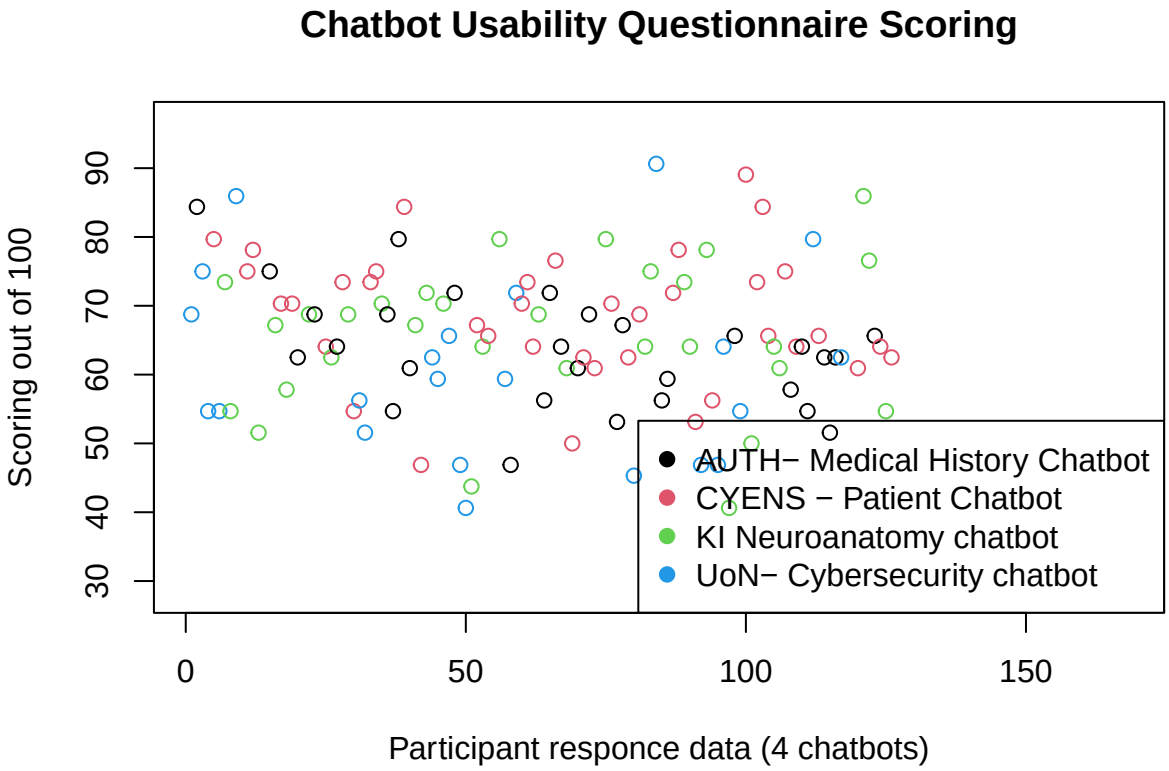


Figure 2.4: CUQ Scatter Plot

- University of Nottingham CUQ score = 68/100

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

Figure (2.4) 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.

2.3 System Usability Scale (SUS) Questions

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

2. Results

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
- I felt confident using the CEPEH chatbot

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	Responses
Agree	66
Disagree	15
Neutral	17
Not Applicable	3
Strongly Agree	23
Strongly Disagree	2

The table **??** 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)	Responses
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 Ease 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 Ease 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

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.

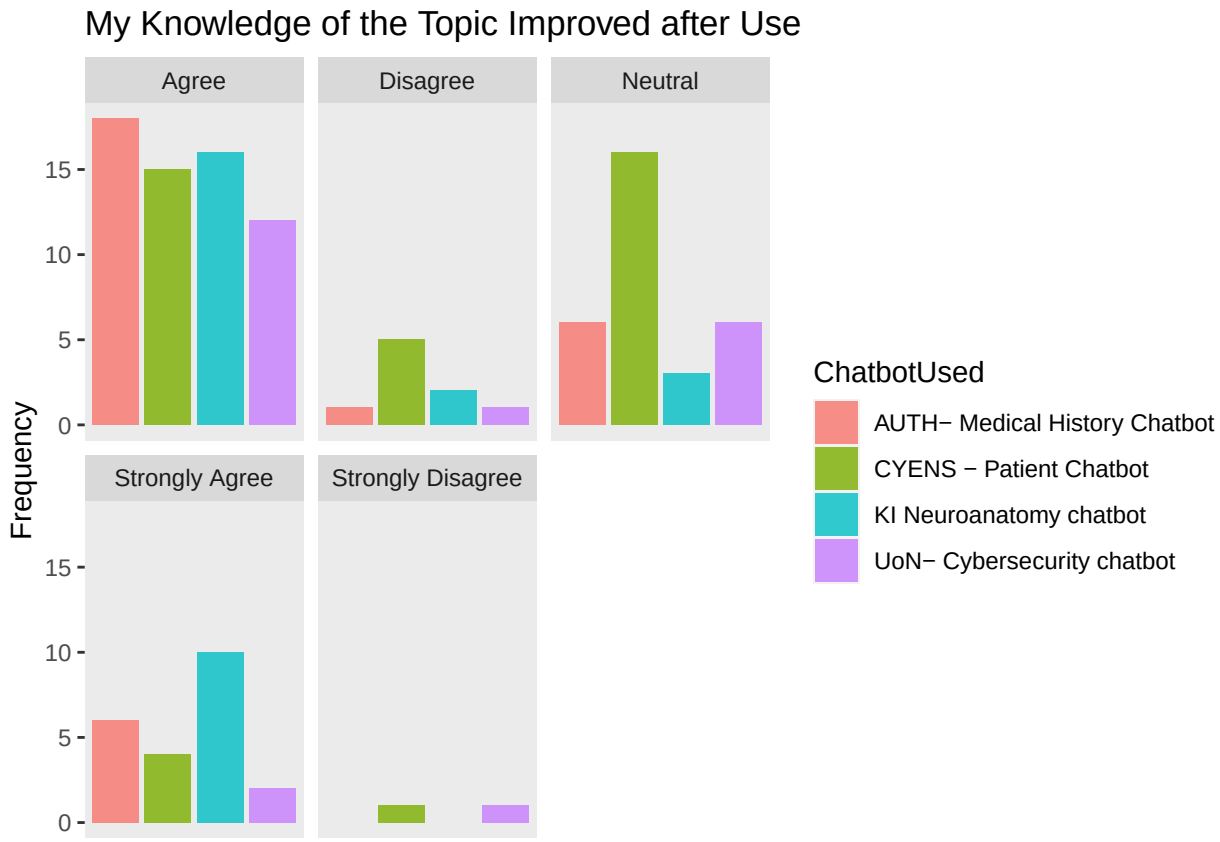


Figure 2.5: Improvements in Knowledge
(#fig:Boxplot knowledge)

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 gure 2.6 shows the ratings by participants of the CEPEH Chatbots to provide them with the necessary course information.

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.

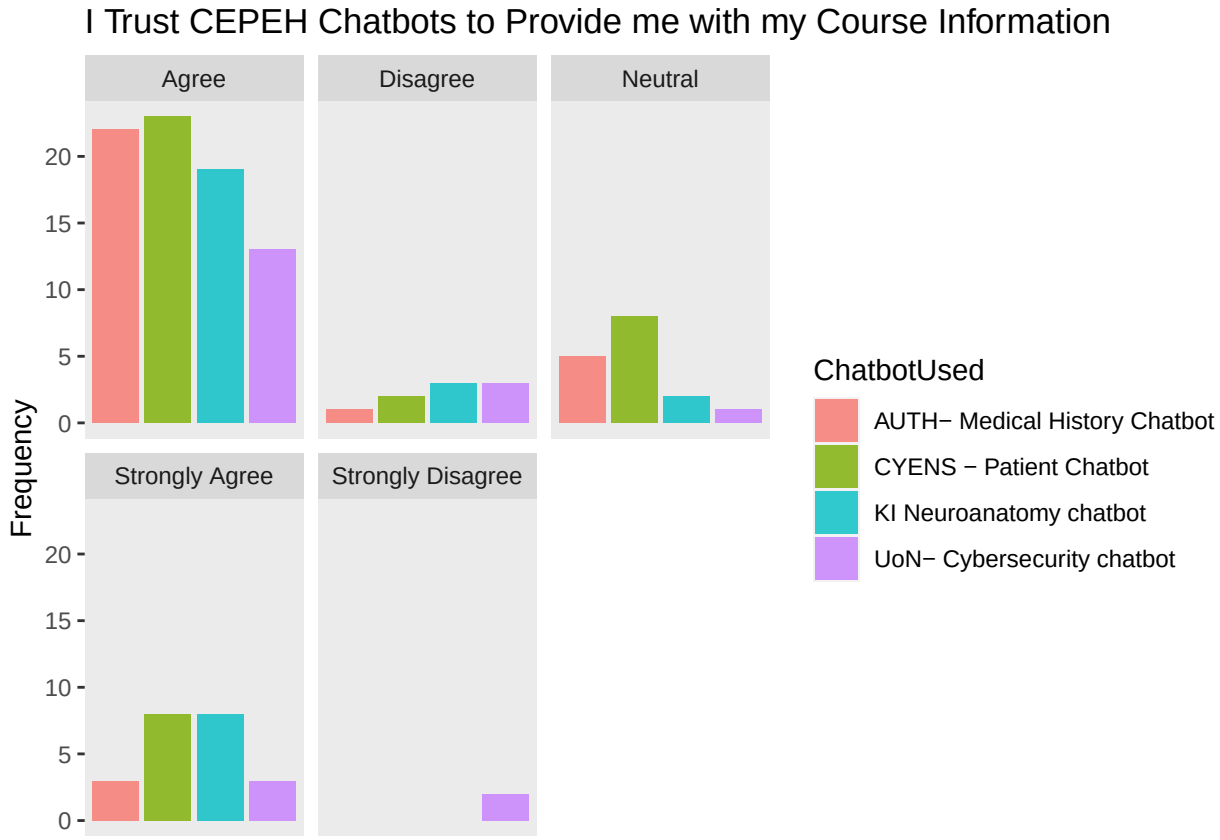
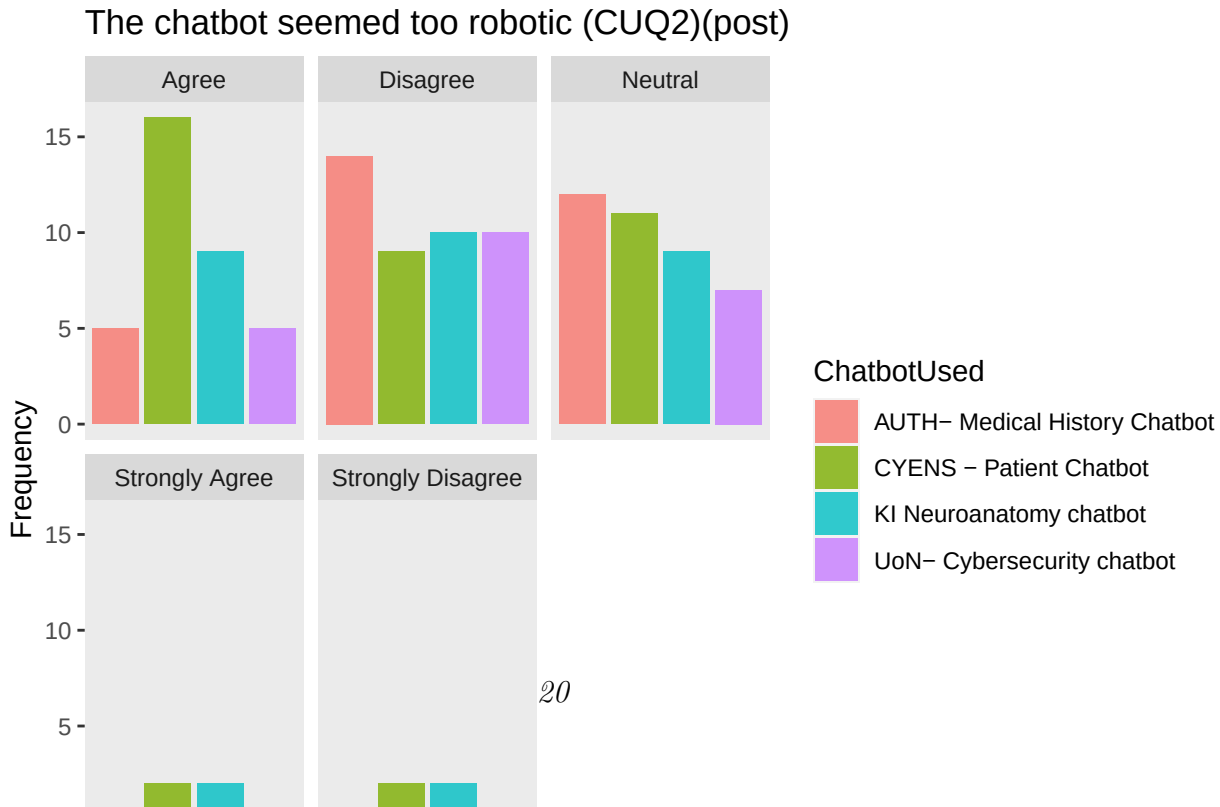
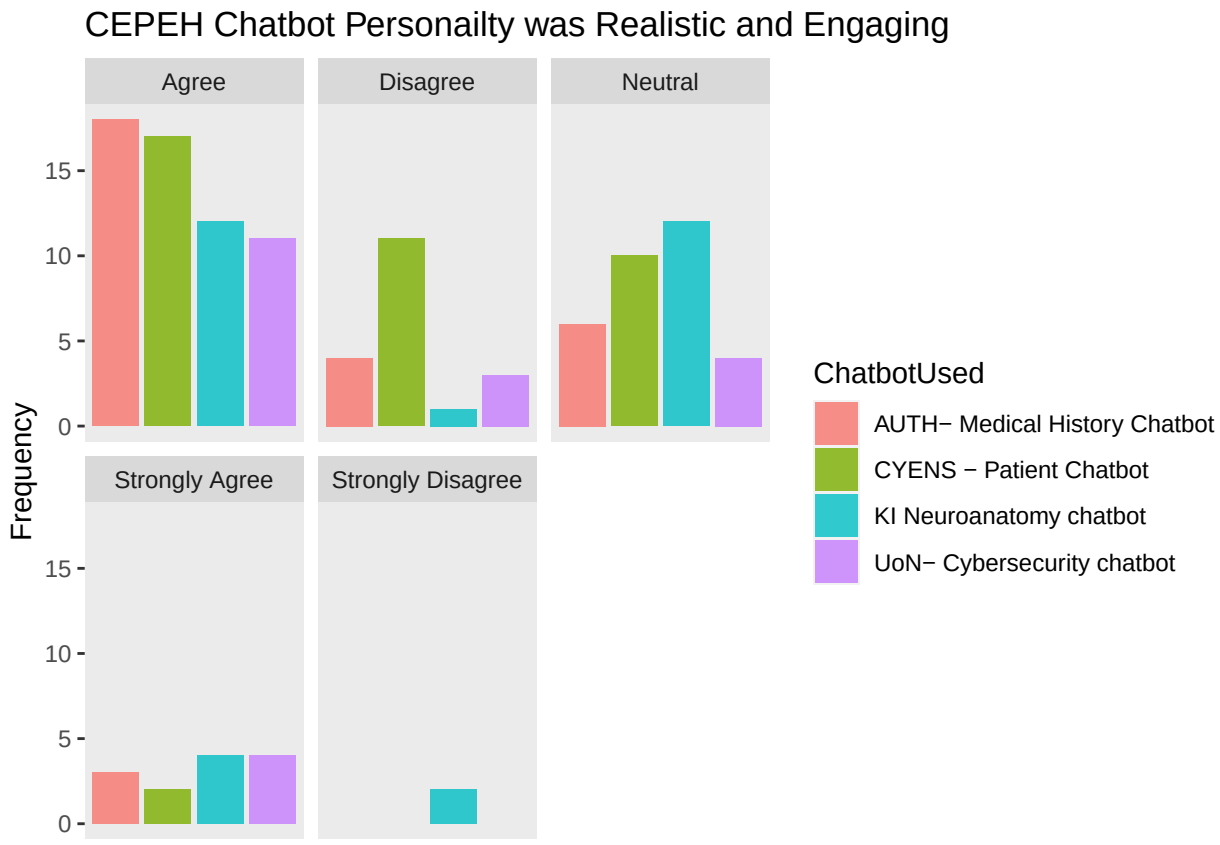


Figure 2.6: Trust Chatbots POST use
(#fig:Boxplot trust)

2.5 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.



There were 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.

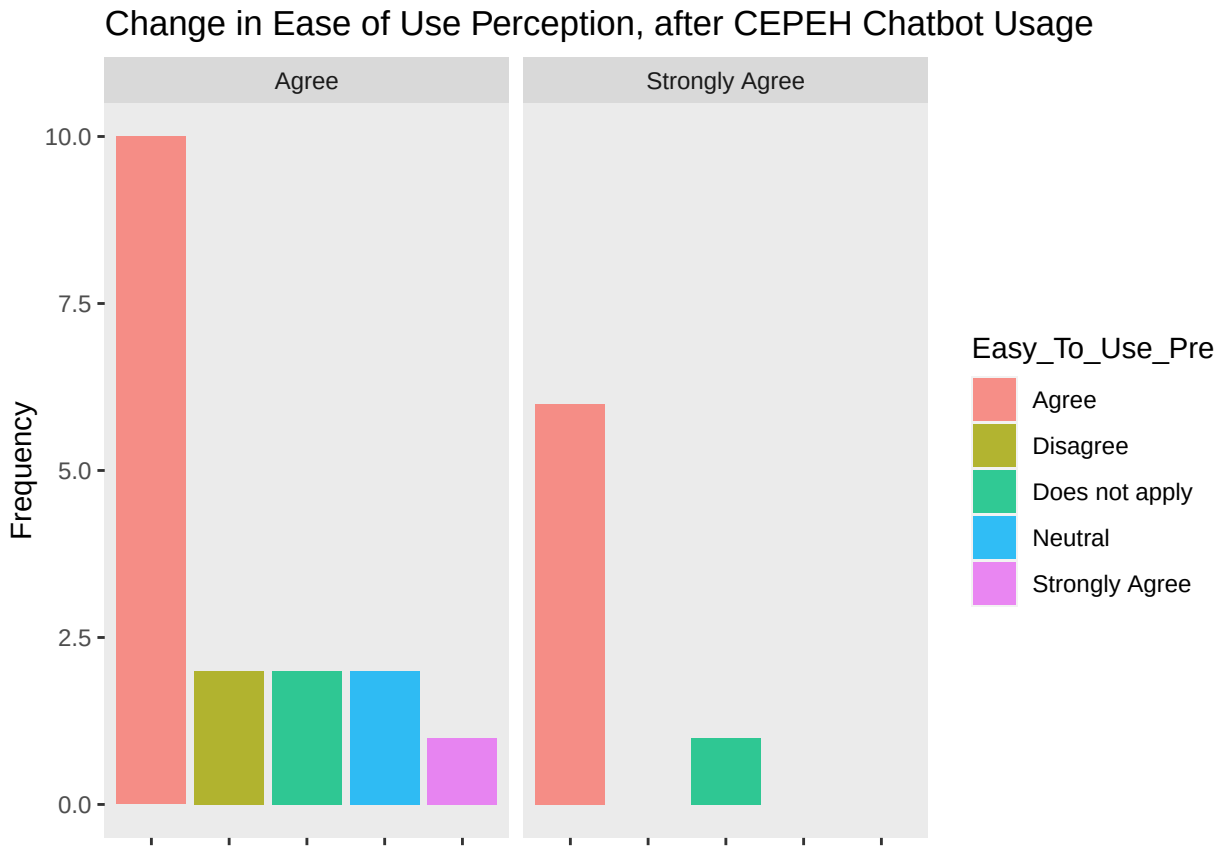


Figure 2.7: Ease of Use Comparison

2.6 Ease of Use and Seeking Support

After usage, there was only agreement in Ease of Use- as shown in (2.7 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.

This rather large table presents all of the descriptive statistic measures AFTER chatbot usage. The Mode is important to show that the majority of participants stated their perceptions, experience, and acceptance increased.

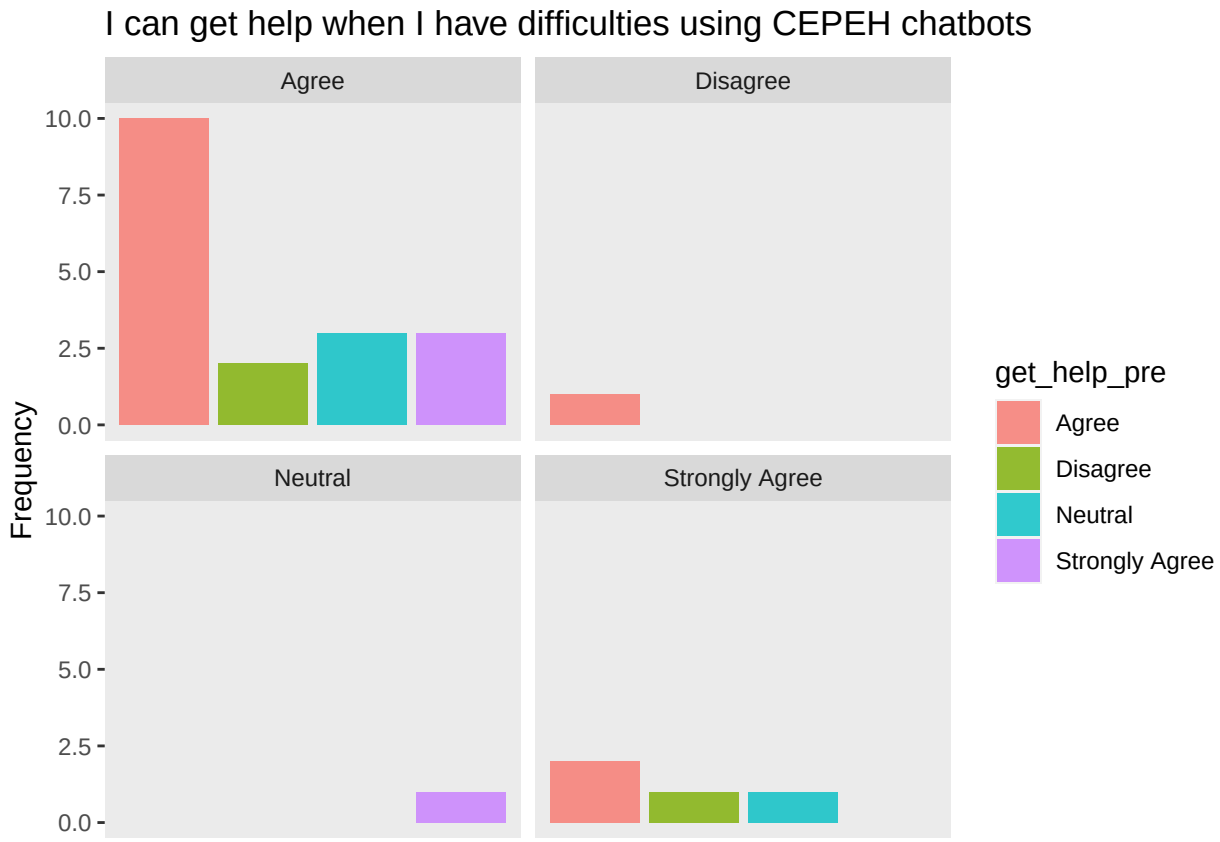


Figure 2.8: Ease of Use Comparison

2.7 Inferential Statistics

Paired t-test involves matching the same participants on a variable before intervention with, ideally, the equal measure of variable after intervention. For this study, we used several metrics from the various questionnaires to facilitate pre-post comparisons. The CUQ was only asked after chatbot usage, however most other questions were able to have a pre-post comparison.

Importantly, there is value in the modes which indicate majority consensus, rather than mean driven t-tests. Being an initial single session with technical orientation for Users as well as practical usage, there is high change that a minority will experience technical or functional problems. Although the majority may benefit, the measures from this minority can significantly impact the measures. For example, if a sample of 42 have the results 5/5 (27), 4/5 (11), and 1/5 (5),

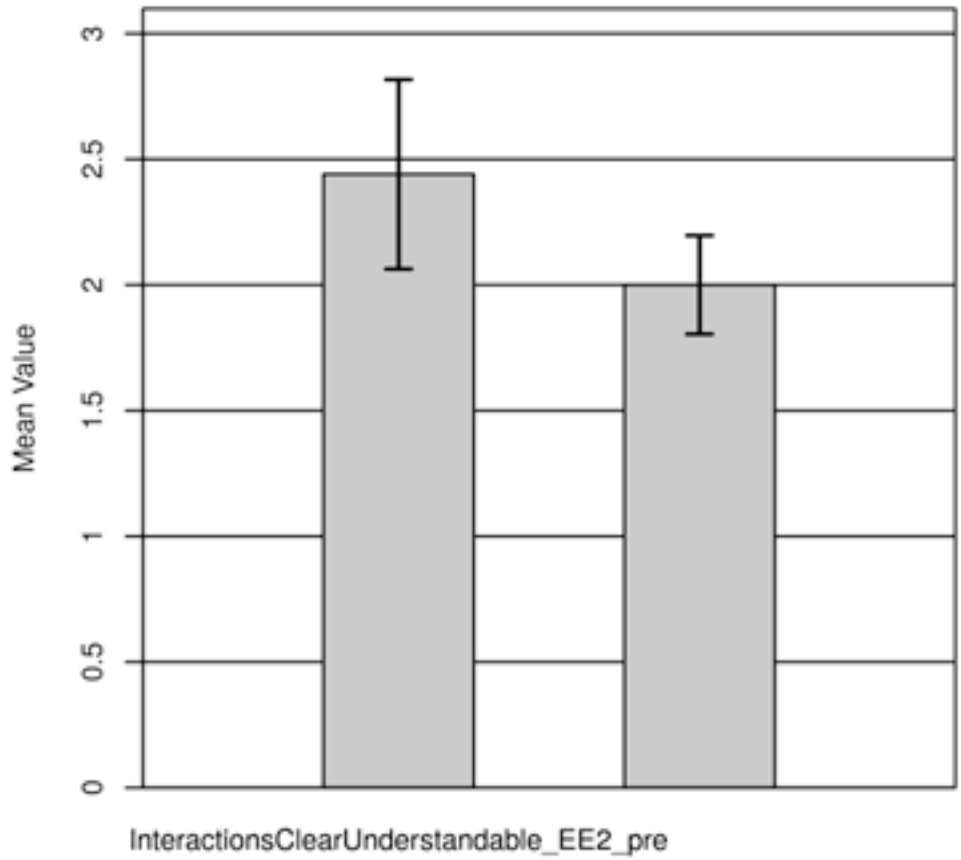


Figure 2.9: Pre-post accomplish quickly

pre		post			Interact Underst
M	*SD*	*M*	*SD*	*T*	*p*
2.44	0.96	2.00	0.50	2.03	.053

Note. N = 25. Degrees of Freedom for the *t*-statistic = 24. *d* represents Cohen's *d*.

Figure 2.10: Table of T-test results

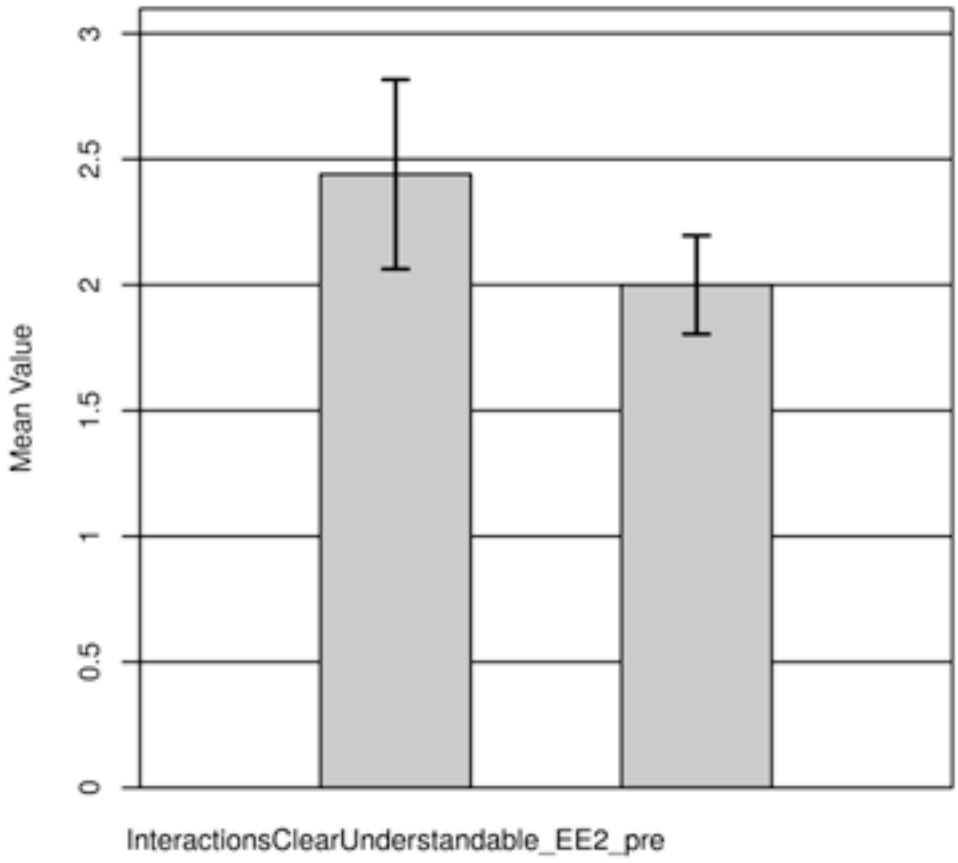


Figure 2.11: pre-post clear

the mode and median are 5 however the mean is 4.38. This can affect parametric or non-parametric results that infer equal experience from participant. We are not overlooking participants but factoring in their experiences as a minority and assessing their results in other ways- i.e., the focus group discussions. This better reflects each participants experience and the accuracy of efficacy of the chatbots. Because of the experimental set-up and the 4 different chatbots, the meaning of the t-tests has small power and effect size. We intended for all metrics to improve, but to have significant findings with the setup does not provide much more information in addition to increased means/scores. We performed paired t-tests, however many paired samples had significant Shapiro-Wilk results indicating the normality assumption were violated. The test is robust to violations, and the

	Mean
Which chatbot(s) did you use during the Training Event	
The chatbot provided the information I needed with minimal commands.	
I felt my knowledge of the topic improved after i had used the Chatbot	
I felt my confidence in understanding the topic improved after I had used the Chatbot	
The chatbot provided me with the type of response i expected from asking a tutor/lecturer	
The information provided was from reliable sources	
I feel the chatbot has a high level of trustworthiness	
The duration of conversations to find my answer, were too long.	
I found the CEPEH chatbots useful in my daily life (PE1)	
Using CEPEH chatbots increases my chances of achieving things that are important to me (PE2)	
Using CEPEH chatbots helps me accomplish things more quickly (PE3)	
Using CEPEH chatbots increases my productivity (PE4)	
Learning how to use CEPEH chatbots is easy for me (EE1)	
My interaction with CEPEH chatbots is clear and understandable (EE2)	
I find CEPEH chatbots easy to use (EE3)	
It is easy for me to become skilful at using CEPEH chatbots (EE4)	
People who are important to me think that I should use CEPEH chatbots (SI1)	
People who influence my behaviour think that I should use CEPEH chatbots (SI2)	
People whose opinions that I value prefer that I use CEPEH chatbots (SI3)	
I have the resources necessary to use CEPEH chatbots (FC1)	
I have the knowledge necessary to use CEPEH chatbots (FC2)	
CEPEH Chatbots are compatible with other technologies I use (FC3)	
I can get help from others when I have difficulties using CEPEH chatbots (FC4)	
Using CEPEH chatbots is enjoyable (HM2)	
I intend to continue using CEPEH chatbots in the future (BI1)	
The videos/images provided were useful to my questions	
The chatbot exceeded my expectation of how it could help me	
The chatbot exceeded my expectation of how it could engage with me	
The chatbot exceeded my expectation of how entertaining it was to use	
I think this learning method could help me to acquire knowledge	
I would be willing to use this learning method again because it has some value to me	

same participants are being tested over Time, rather than different participants. This also indicated that the meaning of the Wilcoxon has limited strength.

2.7.1 Paired sample t-test and Wilcoxon signed rank test

Two-tailed paired t-tests were conducted to examine whether the mean difference is of the following groups were significantly different:

Confidence Daily usefulness Increasing achievements Accomplishing things quickly Increased productivity Ease of use Clear and understandable interactions Use chatbots more frequently

The results showed there were no significant differences in these comparisons. For each comparison the results were:

Confidence- t(24) = -0.35, **p = .731** (prem=1.96, postm=2.04)

Daily usefulness- V = 48.50, z = -0.27, **p = .790** (prem=2.04, postm=2.12)

Increasing achievements- t(24) = -0.18, **p = .857** (prem=2.28, postm=2.32)

Accomplish tasks quickly- V = 36.00, z = -0.25, **p = .805** (prem=2.12, postm=2.16)

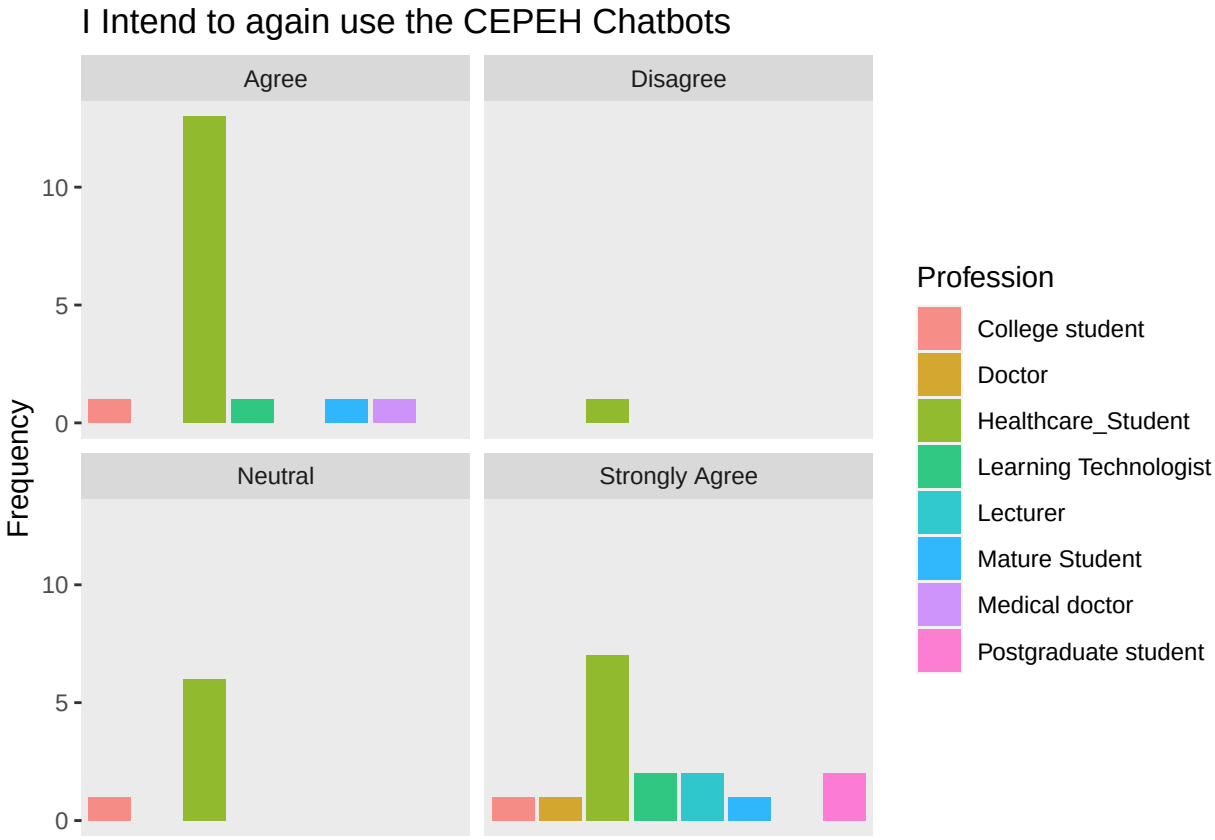
Increased productivity- V = 96.00, z = -1.51, **p = .131** (prem=2.6, postm=2.24)

Ease of use- V = 72.00, z = -1.32, **p = .186** (prem=2.36, postm=2.12)

Clear and understandable interactions- V = 101.50, z = -1.82, **p = .068** (prem=2.2, postm=2.61)

Use chatbots more frequently- t(24) = 0.45, **p = .657** (prem=2.2, postm=2.08)

As predicted, the sensitivity of the t-test meant Wilcoxon test was more appropriate for some measures. The results show minor increases in means for most, but minor decreases in others. These results have high standard deviations which are from a minority of participants scoring low, and explored in the focus group discussions.



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 all 4 chatbots. Further, there were 17 counts of neutral in reuse, which was approximately 4 participants per chatbot (see (2.13).

For CYENS, even though the knowledge of the topic was not perceived to improve by some participants, this box plot shows how 34/42 stated they would reuse the chatbot developed by CYENS.

There was only 1 'Strongly Disagree' response. The agreement options counted for the majority of the data. Repeated Measures t-test, aka paired t-test (before and after measurements)

This t-test compares confident using mobile chatbots before and after CEPEH chatbot usage.

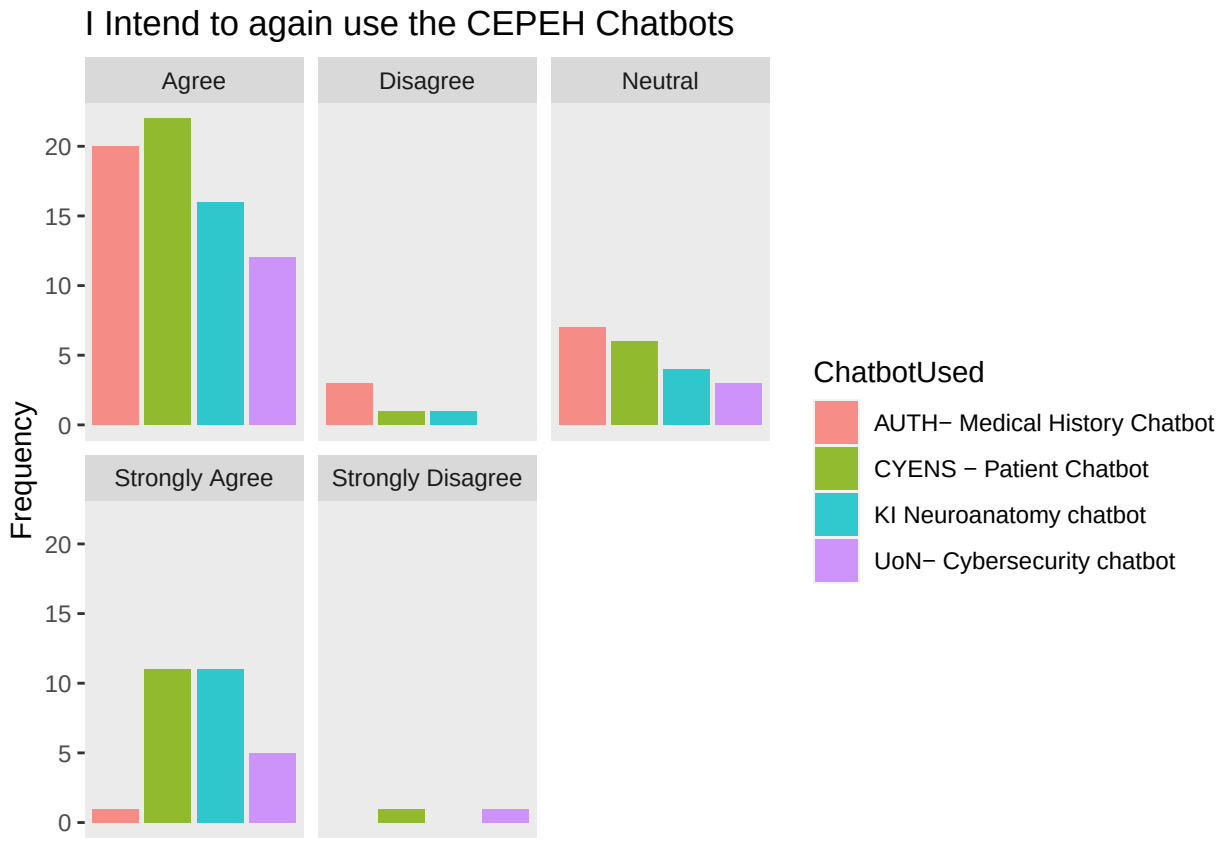


Figure 2.13: Intend to Reuse-Post

output: bookdown::pdf_document2: template: templates/template.tex bookdown::html_document2: default bookdown::word_document2: default documentclass: book #bibliography: [bibliography/references.bib, bibliography/additional-references.bib] editor_options: markdown: wrap: 72 # Text Mining, Natural Language Processing, and Sentiment Analysis —

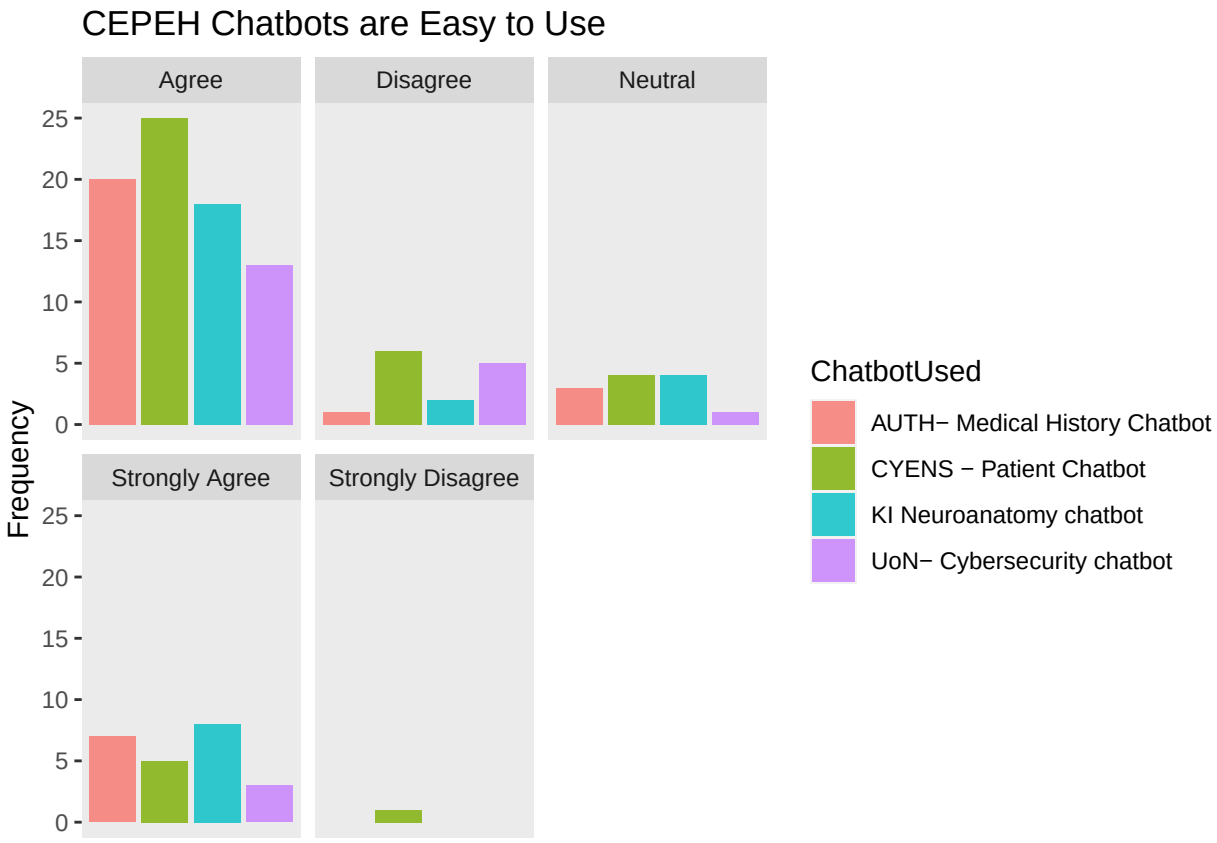


Figure 2.14: Easy to Use- Post

3

CEPEH Focus Group Discussion Analysis

The focus group discussions provided a lot of feedback for how the participants experienced their interactions with the chatbots, and how the CEPEH team can improve them, improve the design and development processes, and improve uptake and sharing.

One method of analysing this data is with use of text mining and data manipulation, creating word clouds, sentiment analysis, and using a model which can distinguish the unique themes in text, and highlights for us what text is used to create these themes.

Therefore, we have created a model to allow efficient and intelligent analysis of this open/free focus group data.

3.1 Tokenising

Firstly, we tokenised the words from the FGDs. A Token is "a meaningful unit of text, most often a word, that we are interested in using for further analysis". For each word we give it a property that we can call upon later.

The data manipulation for this included removing punctuation, converting to lower-case, and setting word type to word (and not such types as "characters", "ngrams", "sentences", "lines" etc)

3.1.1 Stop words

The model then removed words with meaningless function. These are called stop words. Words like "the", "of" and "to" are the most frequent words found, technically, but are of little interest to us.

We also created a custom list of stop words for CEPEH. We know participants may mention other objects, and the list was as followed: found; chatbot; chatbots; presentation.

The data was ready for analysis by the model. We ordered it to find the most frequent words. Below is a table with the 6 frequently occurring words, showing how Stop words have now been filtered.

word	n
information	11
helpful	8
understand	8
idea	7
ideas	7
lot	7

This word list can then be used for sentiment analysis, (see *Sentiment Analysis* section), in addition to frequency of words.

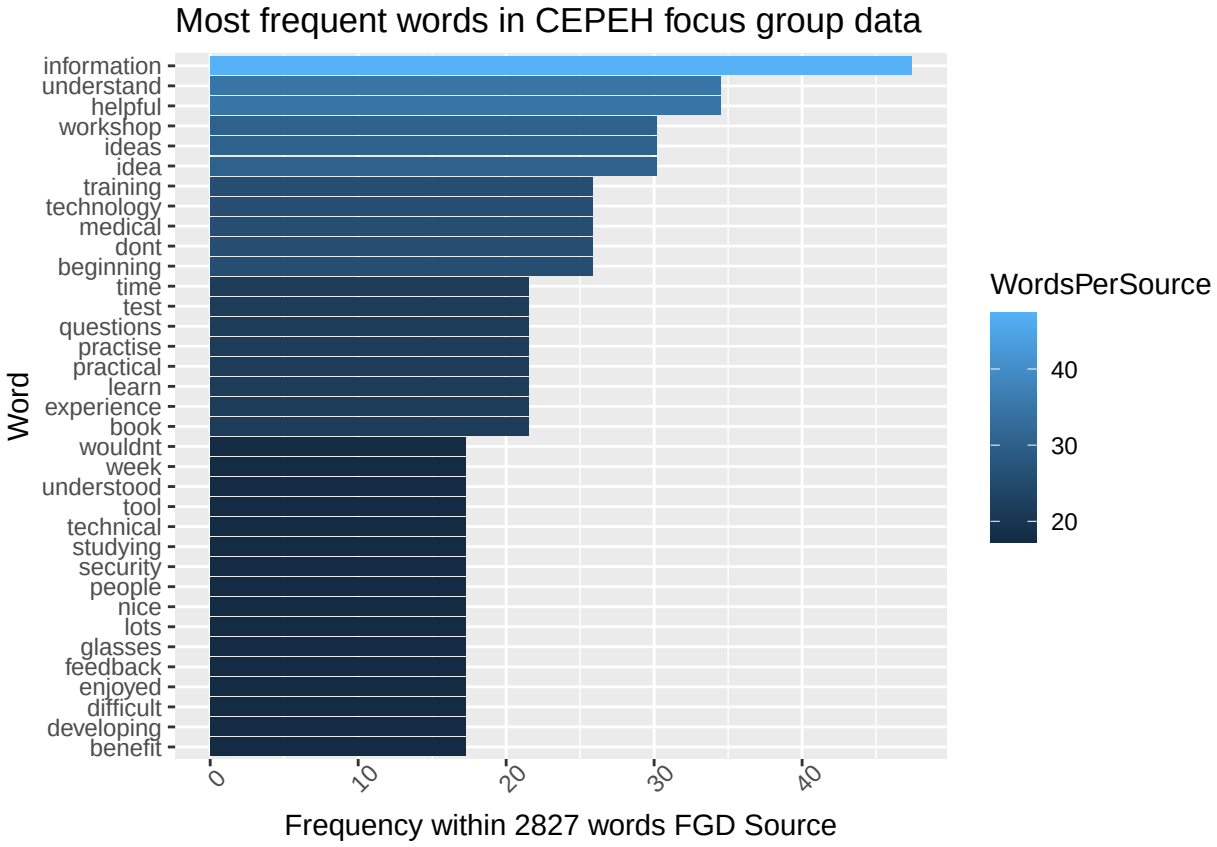
3.2 Plotting word frequencies - bar graphs

3.2.1 Normalised frequency

With this information a list of top words from the participants in the FGD can be rendered and after some modifications, a graph of the top 20 words is produced, with better aesthetics. This is a better way to understand this data, and the axis can be normalised for the frequency of occurrences in accordance with the source text. The raw text had 2827 words in total. Therefore we can mutate the ratios to reflect this.

3.2.2 Plotting normalised frequency

Now we can plot, for example, the 20 most frequent words when normalised by the source text.



In summary, this understanding of frequent words can help to understand common concurrences and extrapolate to a larger audience. If scope and impact of CEPEH chatbots increased we can understand the type of themes and trends may occur, based on such FGD analysis.

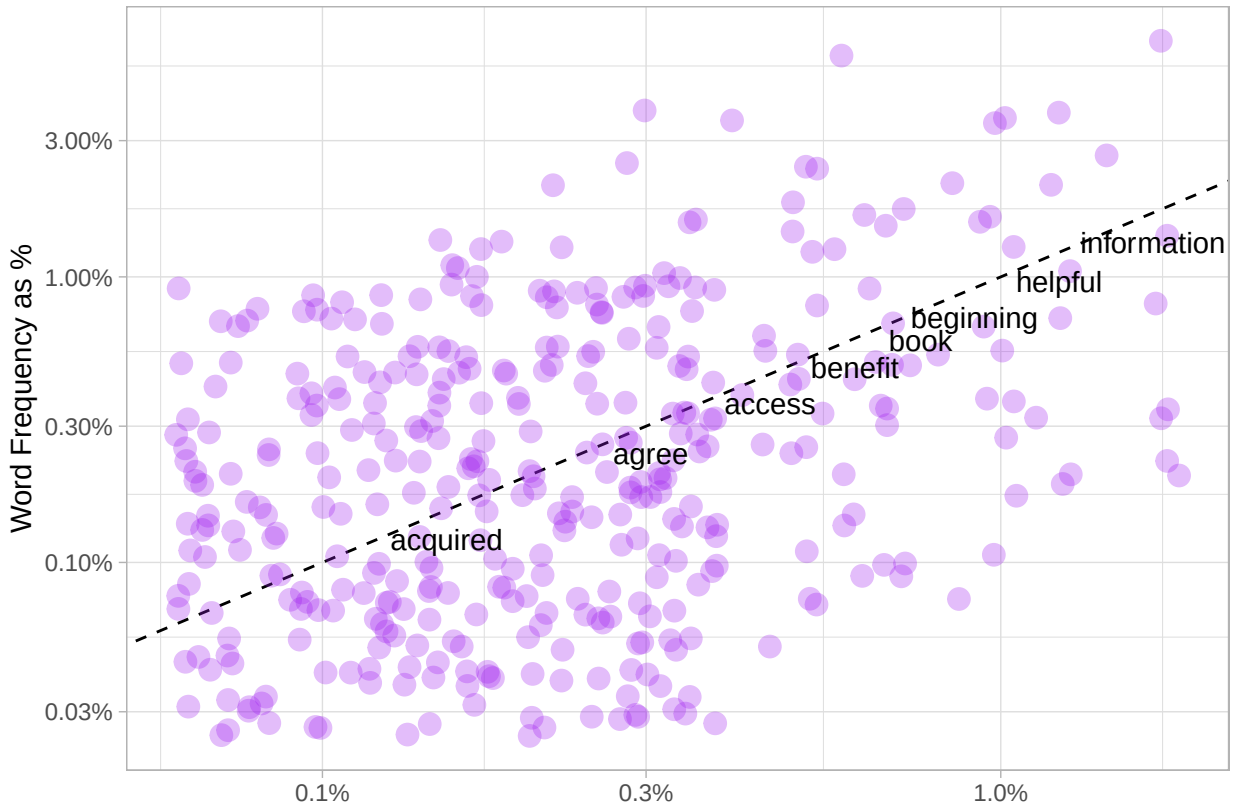
3.3 Word clouds

To visualise the most frequent words in another format, below is a word cloud which presents the word size to indicate the frequency- words that occur more often being displayed in a larger font size. This has a normalised data frequency in accordance to the FGD source document analysed.

We understand the context has been reduced for each word. However, in general there can be categorised positive/negative words from the word cloud: Positive words are- benefit, practical, nice, helpful, learn, ideas, and enjoyed Negative words are- difficult, test (who likes a test?), don't, and 'lot' may be negative if there is a 'lot' of information.

3.3.1 The vocabulary of Texts

Here is a graph that has plotted the words in places depending on the word frequencies. Additionally, colour hotspots shows how different the frequencies are - darker items are more similar in terms of their frequencies, lighter-coloured ones more frequent in one text compared to the other.



3.4 Sentiment analysis

What is the sentiment of all participants? What is types of emotional words are being used? The preparation of these words has some use in understanding the frequencies, but their emotional valence are not compared. The table above has the word *'helpful'* which has a positive connotation, however there are 386 words, with many having several occurrences.

As the table below shows. the FGD data has been analysed for sentiment of each word, and has been calculated to have 62 positive emotional valence of words, with 24 negative valence of words. These are from a **Bing sentiment lexicon** which is the most popular English language dictionary.

negative	positive	total_score
24	62	38

Unfortunately, there is little research using sentiment analysis for chatbot related focus group results that can help to understand the scoring found. However,

on a basic interpretation the higher the score the better the chatbots were discussed in the FGD's. A score of 72% (62/(24+62))) would be in 3/4th quartile in distribution of sentiment distribution. Alternatively, $62/24 = 2.58$ would state the ratio that for every 1 negative word recorded, there were 2.58 positive words recorded.

4

Discussion

Contents

##Summary of Findings

The Chatbots were beneficial for the majority of users, on metrics which are indicative of uptake. Learners have lots of other choices such as YouTube, but there is a certain need for personalised information gathering, this can save time and prevent learning incorrect information. This was one reason why they were rated positive as they are able to streamline data finding for learners in a format that is understandable and easy to them.

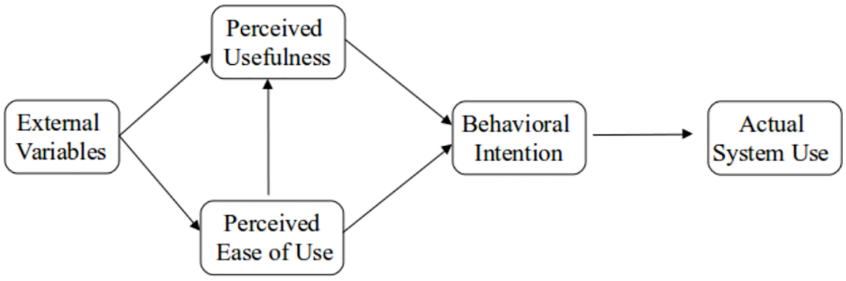


Figure 4.1: TAM Model processes

4.1 Quantatative Results

4.1.1 CUQ

Holmes et al. [https://dl.acm.org/doi/10.1145/3335082.3335094] designed the CUQ to be comparable with the system usability scale (SUS).

We have calculated both these scores out of 100 to allow the same benchmark, which is 68. A score of 68 is at the centre of the range is thought of as "C". The average benchmark for CUQ is 68, and in initial/pilot studies 68 may be considered higher than expected when considering technical issues, less developed user interfaces etc.

Previous studies evaluating chatbots have had similar score. For example, in 2022 Link found a physical activity promotion chatbot received 64.5/100, with lowest score at 40.6

4.1.2 TAM

4.2 Qualatative Results

4.3 Limitations

4.4 Conclusions

This document details the evaluation of 4 Chatbots created using the ASPIRE model for healthcare pedagogy. The feasibility and acceptance from the end-users were the focus, to understand future uptake,impact, etc. Identifying the feasibility

of such resources into formal training is essential and studies have promising results when following structured design principals. 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.

Those results can be interpreted that the learning objectives of the training event was chosen appropriately for the diverse audience including clinicians, academics, researchers, and learning technologists/IT specialist resulting to a successful training event that enable participants to take the acquired knowledge back to their organisations in order to co-design and implement. As it was expected and can be depicted from self-confidence statements that some participants being very confident before the event, not all the objectives expected to be reached by everyone, since the training was targeting both technical and non-technical participants. However, on both average and individual matched responses participants self-statements showed that they improved their knowledge and understanding in using co-creation approaches to develop digital education resources and in designing and developing chatbots as educational resources.

output: #bookdown::html_document2: default #bookdown::word_document2: default bookdown::pdf_document2: template: templates/template.tex documentclass: book

5

(Additional Analyses) Training Events

Contents

5.1 CEPEH Training Event C1

The CEPEH training event C1 held at the premises of University of Nottingham aiming to prepare participants for the practical elements of co-creation and implementation of chatbots as an educational resource. It combined both theoretical and hands-on training. 15 participants were from RISE, AUTH, UoN.

Project managers of partners signposted the person involved, and relevant announcements were made though social media channels to the wider public. External to the project speakers were from University of Leeds, and Computer Science Department of University of Nottingham. It included academics, medical doctors, and researchers with focus both on clinical research and digital innovations in

healthcare education and IT specialist/learning technologists 11.18 years of experiences (SD=7.2). A balance between male and female participants achieved.

Participants were asked to highlight what they liked for each day and how each day can be improved. Findings are described below per day of the training event

Day 1

The participants comment that they liked the design method for educational resources presented using a co-creation approach, they liked the interactions with other groups, and they liked the overview of existing chatbot resources of the partners. On the areas that can be improved, more media material were requested.

Day 2 Participants enjoyed the presentation from the invited speaker from another faculty of the University of Nottingham, the CEPEH recources presented and the storyboarding process. Participants highlighted that the participation of more clinicians in the event would be an added value in regards with the storyboarding process.

Day3 Participants liked the hands-on activities of the day also enjoyed the creativity of the groups on the online chatbot development tool. As an area of improvement, participants wanted more time on hands on sections.

5.2 CEPEH Training Event 2

Pre-Training Event survey May 9th-13th 2022 Thessaloniki, Greece

Twenty-six participants attended the Training Event, along with approximately 10 staff members. There were 21 undergraduate students and 5 postgraduate students, who completed the survey for a total of 26 responses. There were 86% of participants who stated they had not been to a similar event like the training event CEPEH facilitated. There were 90% of students who found the event schedule very organised, and 70% agreed most of the planned sessions were relevant to that interest with the remaining 30% not having enough experience to understand the context to determine if they are interested in the training event. There were 95%

of students agreeing or strongly agreeing the training event location is great, the remaining person did not leave additional comments.

Table 1 suggested attendees had minimal intention to share their own ideas due to lack of previous experience of attending such events, or due to lack of knowledge on the area. However, most were interested in listening to other groups and hearing contextual cases in healthcare.

There were 77% of participants stated they were novices in experience with chatbots in healthcare and were attending to learn more. The remaining 23% (7 students) stated they were competent and had limited experience with chatbots in healthcare.

One day had several events regarding cybersecurity in healthcare. When asked before these events, 83% stated they were neutral or disagreed that they felt confident about their cybersecurity knowledge in healthcare. In addition, 80% stated they when neutral or disagreed that they felt they had strong cybersecurity safety in healthcare. Table 2 shows the main pre and post results suggesting a positive experience for more than 75% of attendees on all measures.

There were 90% (23) of students who heard about the event through a lecturer or a professor, the CEPEH newsletter (2), and 1 person was informed through the anatomy tutoring system at Karolinska Institute. Additionally, 60% suggested the training event to somebody else before the course started.

There were six individuals who stated neutral or disagree when asked if having issues on registration or finding the information for the event. This may have been due to being dependent on emails to receive the information, instead of a dedicated website where the information is available anytime.

As this was face-to-face, participants were asked about sufficient Covid-19 precautions in place at the facility, 94% agreed with sufficient precautions, two individuals stated no but did not give further information in the additional input box provided. In summary, most participants were undergraduate students with novice experience, happy with the training event location, felt the sessions were relevant to them, and most shared the event with their colleagues. The values of co-creation,

chatbots in healthcare, and taking patient history were bestowed to students in an engaging and well-received manner. Notably, the highest ratings were for staff friendliness which is key to engagement and consistent interaction throughout the intense and long 5-day duration. The sessions were recorded there for the online recordings may be viewed with higher numbers over the subsequent weeks.

5.3 # bibliography: [bibliography/references.bib, bibliography/additional-references.bib]

Appendix

Appendices

A

The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readability:

In 02-rmd-basics-code.Rmd

And here's another one from the same chapter, i.e. Chapter ??:

References

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life.* John Murray.

Goethe, J. W. von. (1829). *Wilhelm Meisters Wanderjahre oder die Entsagenden.* Cotta.