1 Results

1.1 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 1.1: 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.

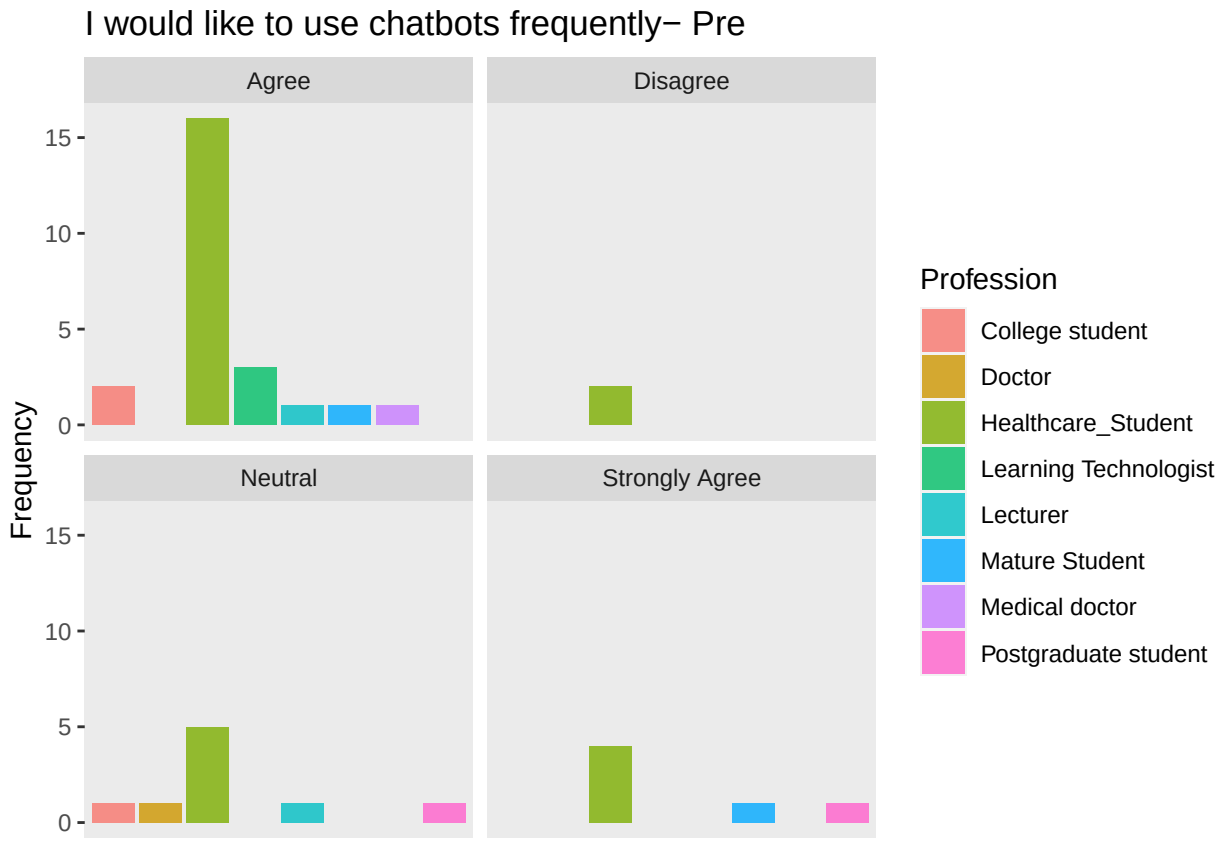


Figure 1.1: Chatbot Usage History- Pre

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 (1.1) shows learners perceptions of ease of use of mobile app and other educational mobile resources

(1.2) 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:

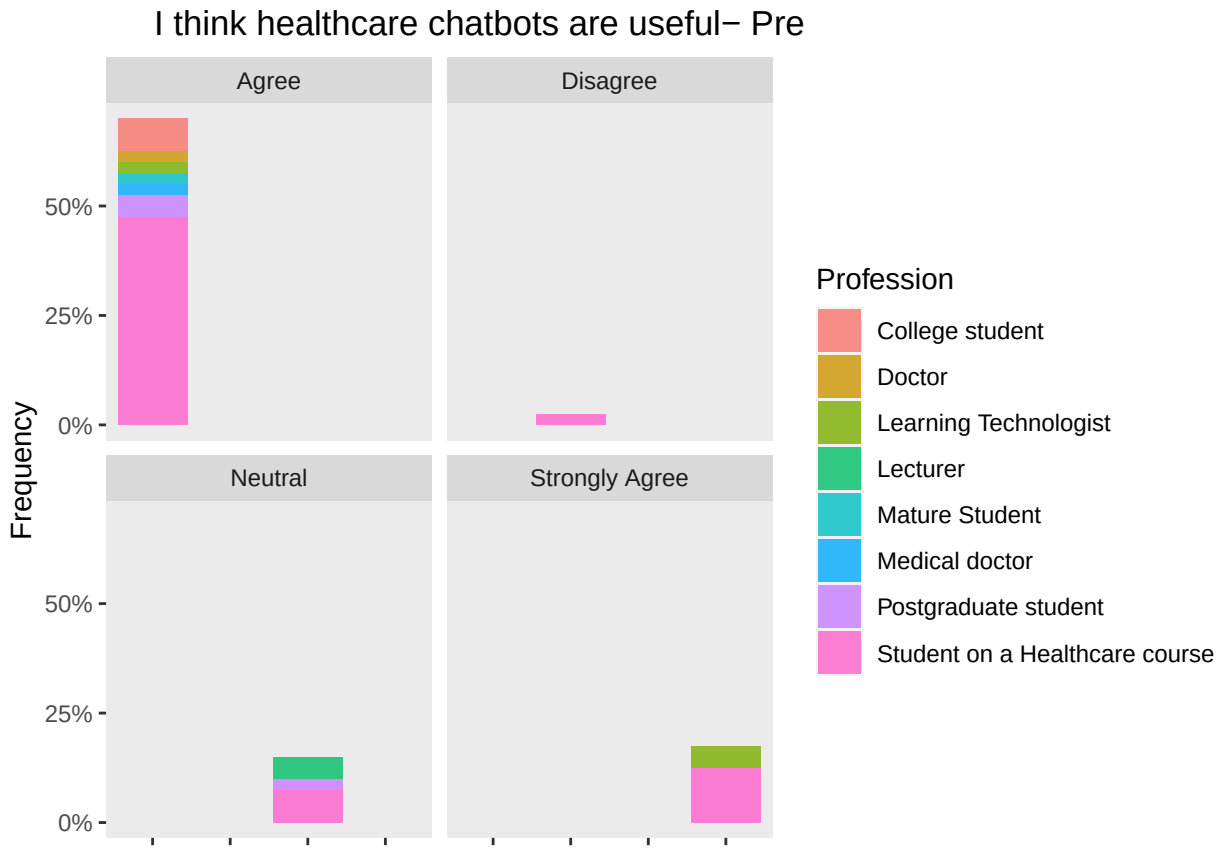


Figure 1.2: Chatbots are Useful Opinion- Pre

1.2 Chatbot Usability Questionnaire (CUQ)

1.2.1 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:

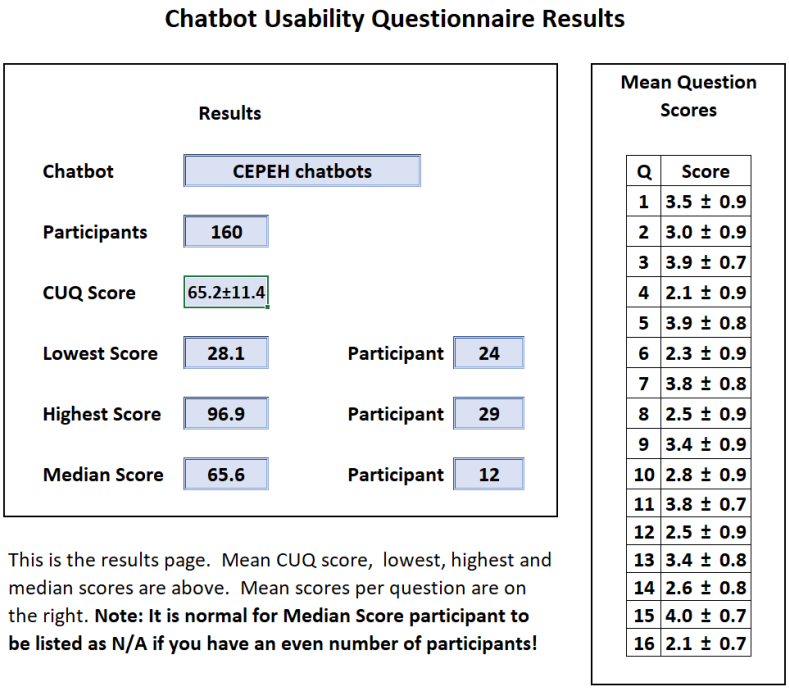


Figure 1.3: CUQ CEPEH Score

- 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 (1.4) 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.

1.3 System Usability Scale (SUS) Questions

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

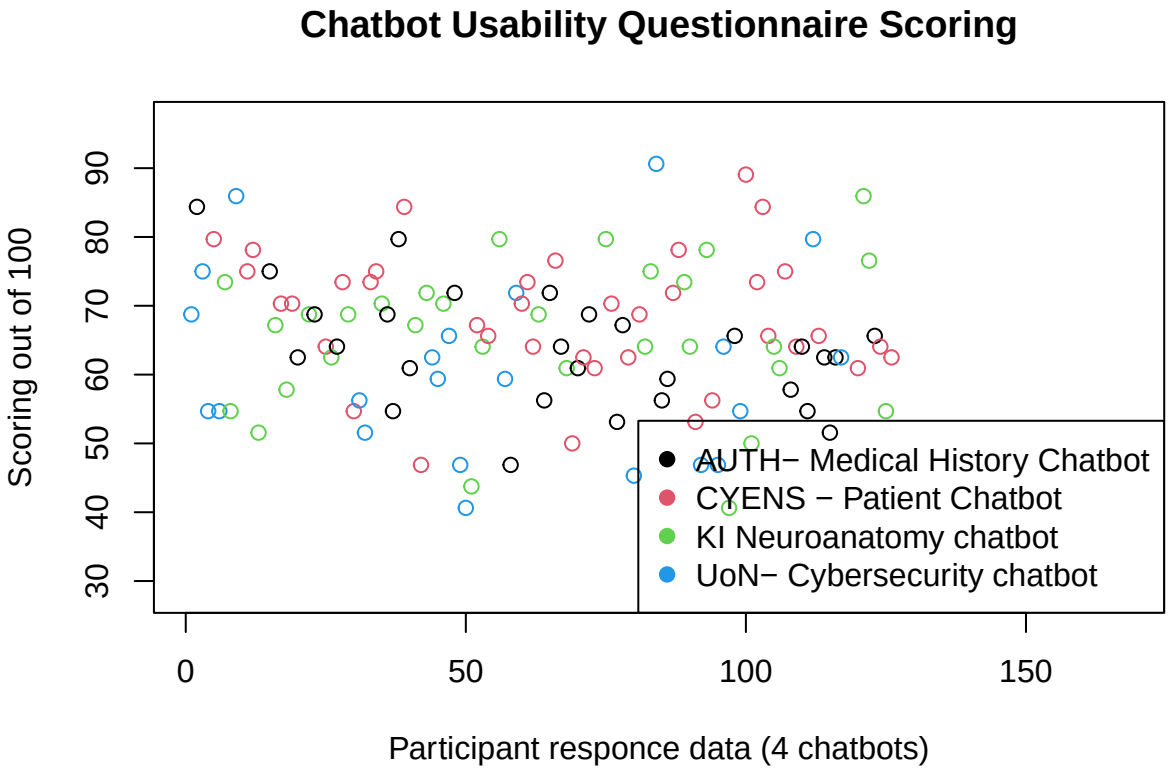


Figure 1.4: CUQ Scatter Plot

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

Keep Using CEPEH Chatbot	Responses
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.

1.4 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.

1.4.1 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 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.

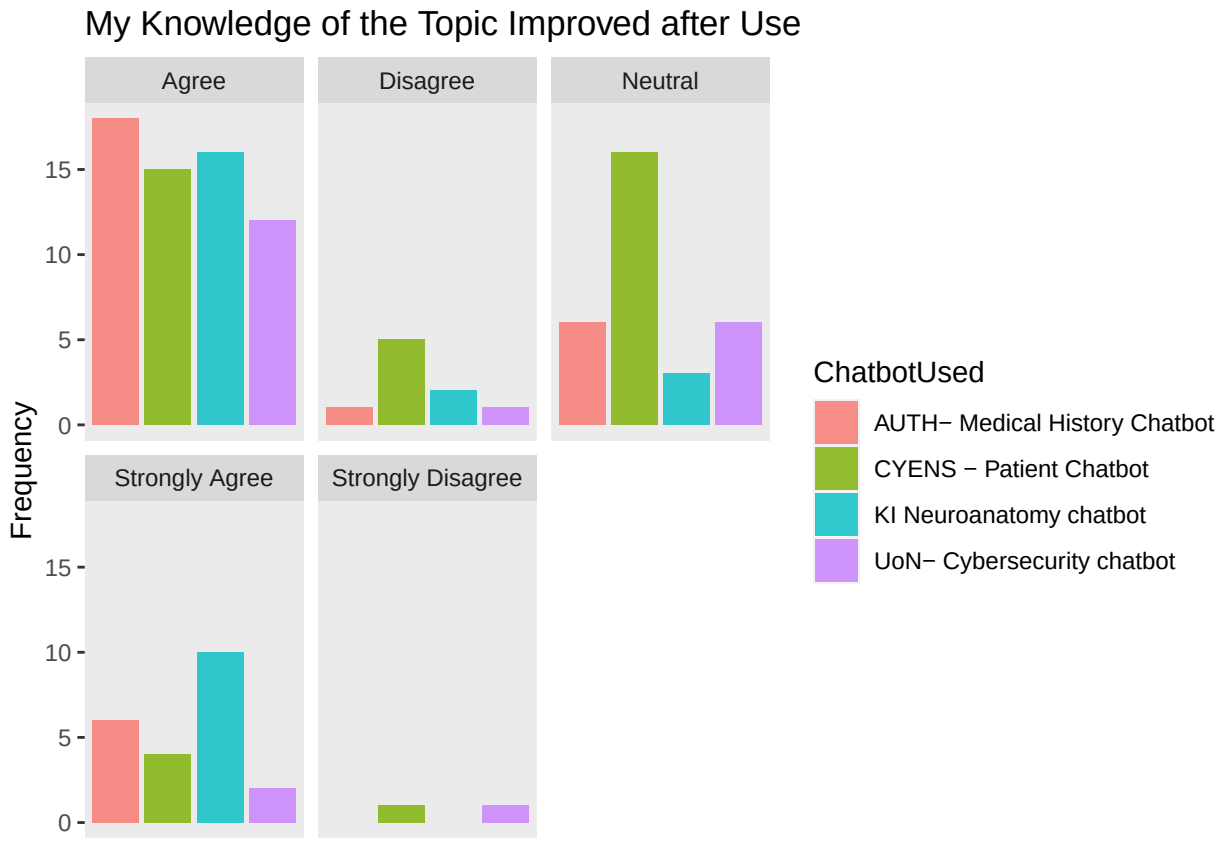
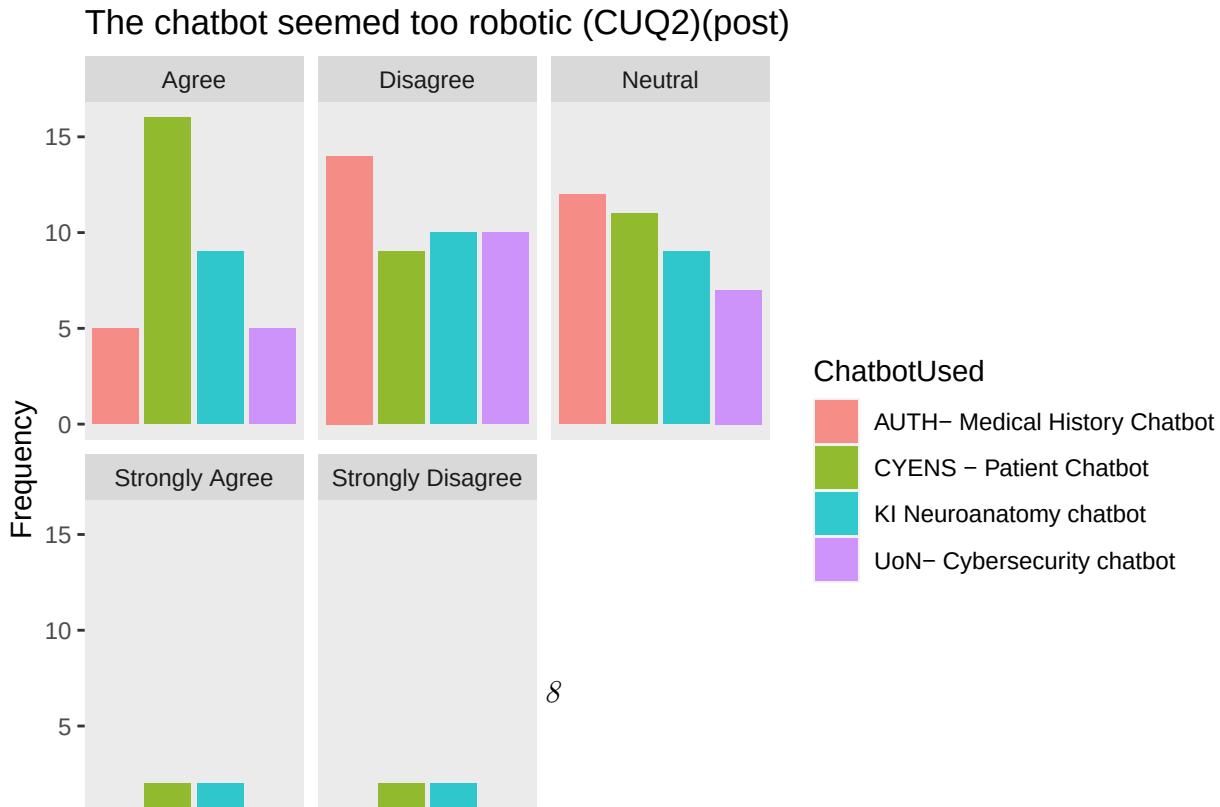


Figure 1.5: Improvements in Knowledge
(#fig:Boxplot knowledge)

1.5 Personality and Interactions



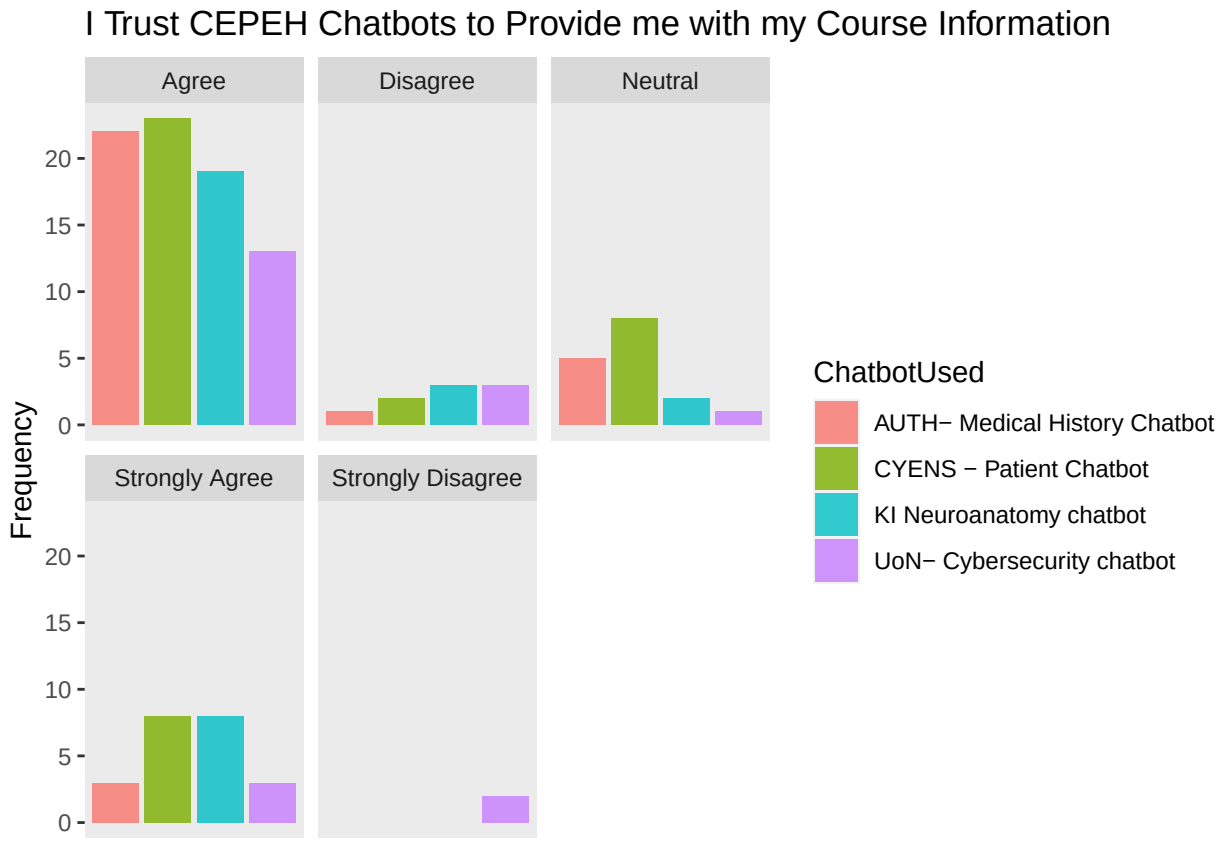
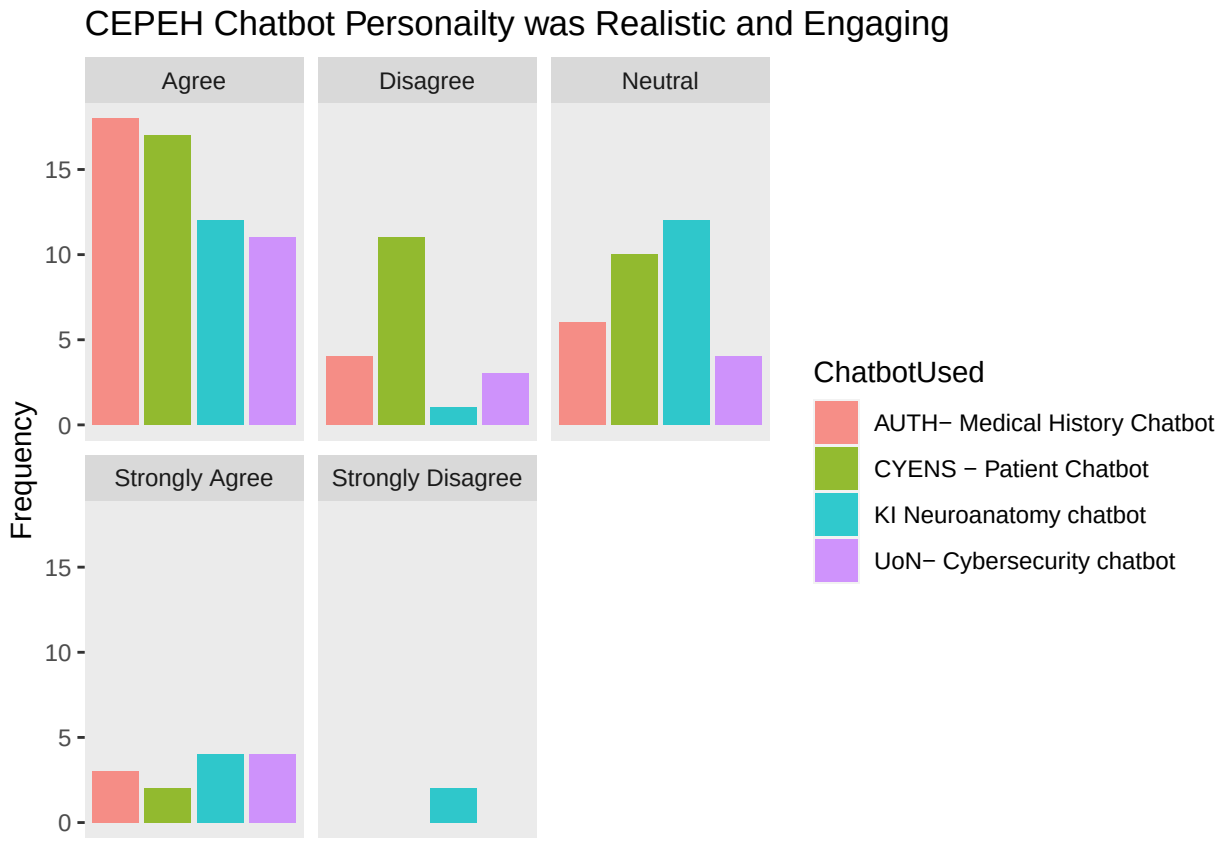


Figure 1.6: Trust Chatbots POST use
(#fig:Boxplot trust)

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.

1.5.1 Ease of Use and Seeking Support

After usage, there was only agreement in Ease of Use- as shown in (1.7 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.

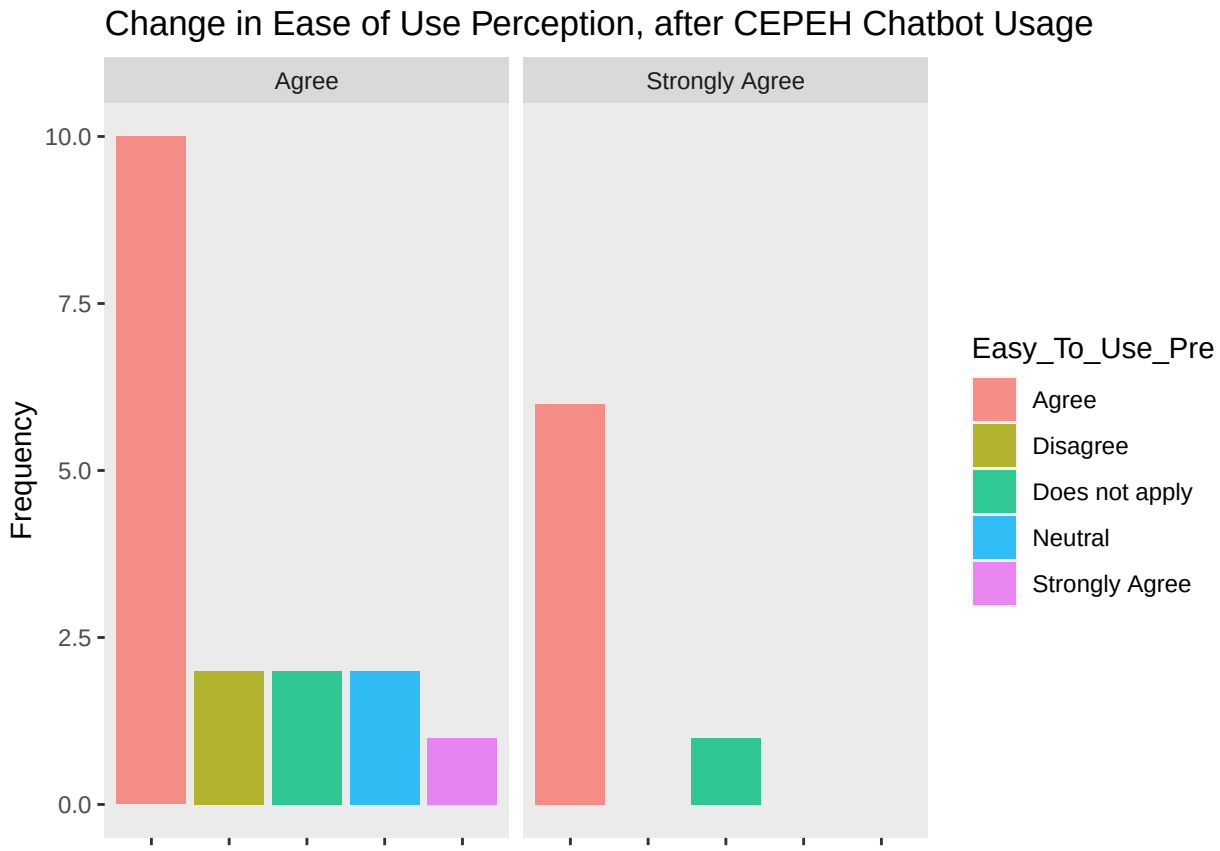


Figure 1.7: Ease of Use Comparison

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.

1.6 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

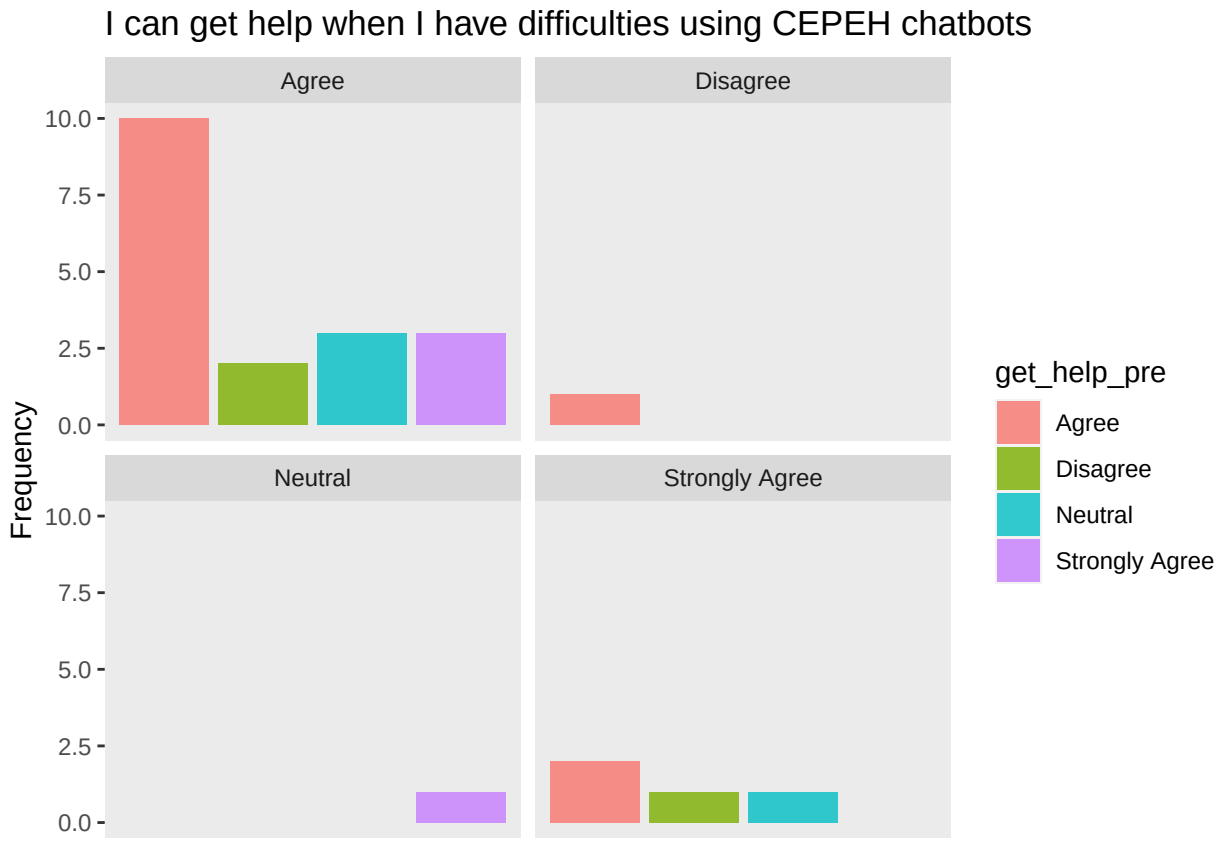


Figure 1.8: Ease of Use Comparison

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), 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

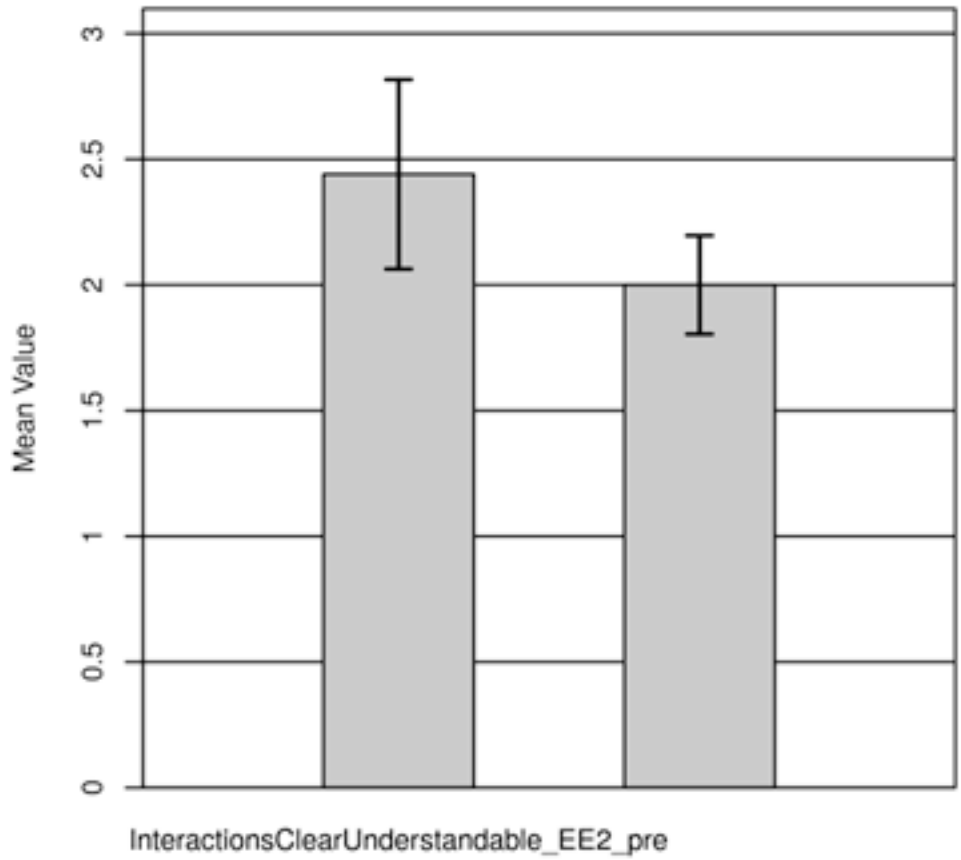


Figure 1.9: Pre-post accomplish quickly

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 1.10: Table of T-test results

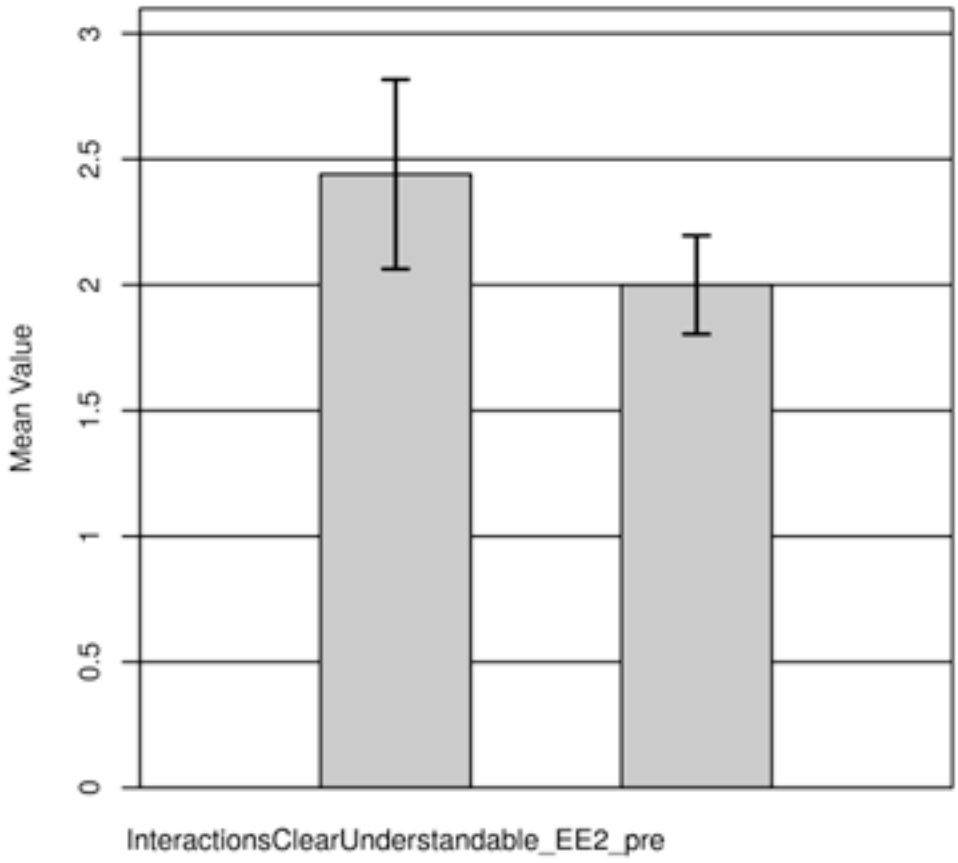


Figure 1.11: pre-post clear

normality assumption were violated. The test is robust to violations, and the same participants are being tested over Time, rather than different participants. This also indicated that the meaning of the Wilcoxon has limited strength.

1.6.1 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

	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	

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 some, 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 (1.13).

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.

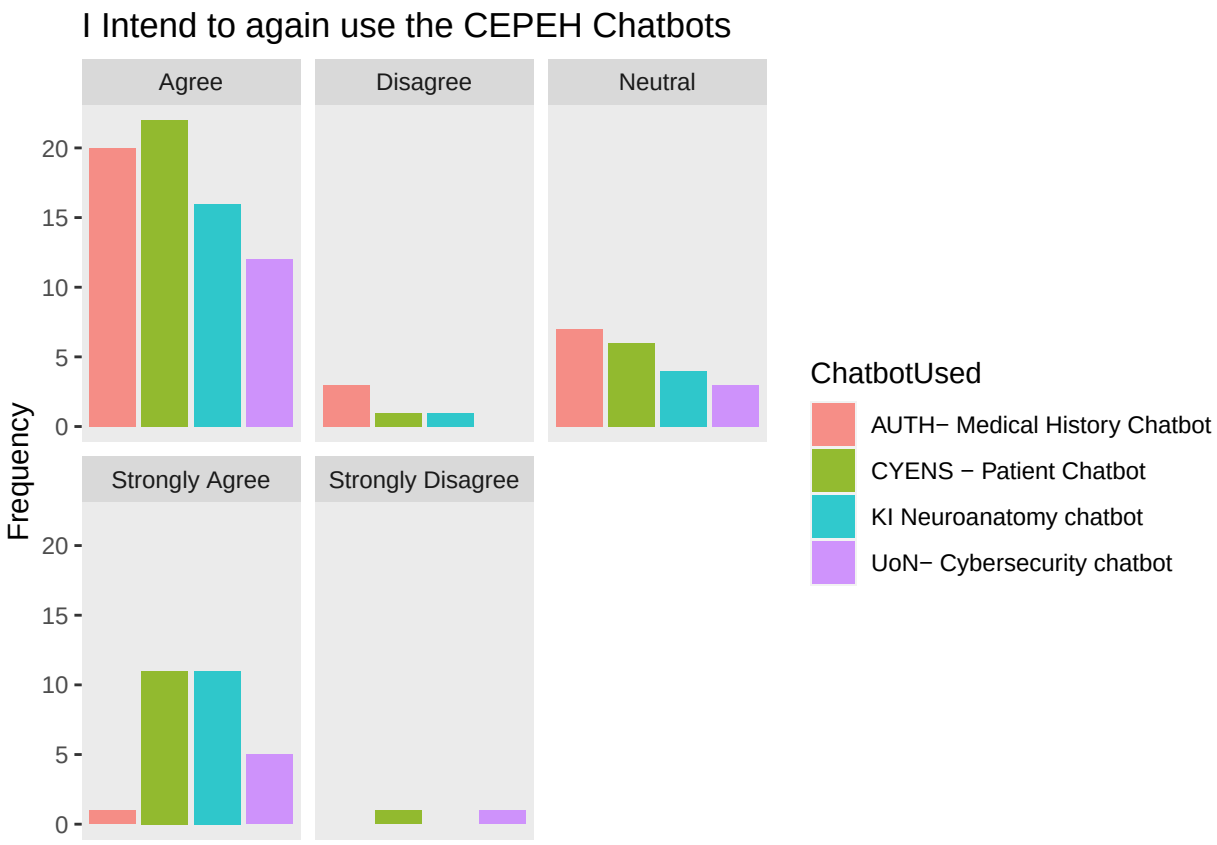


Figure 1.13: Intend to Reuse-Post

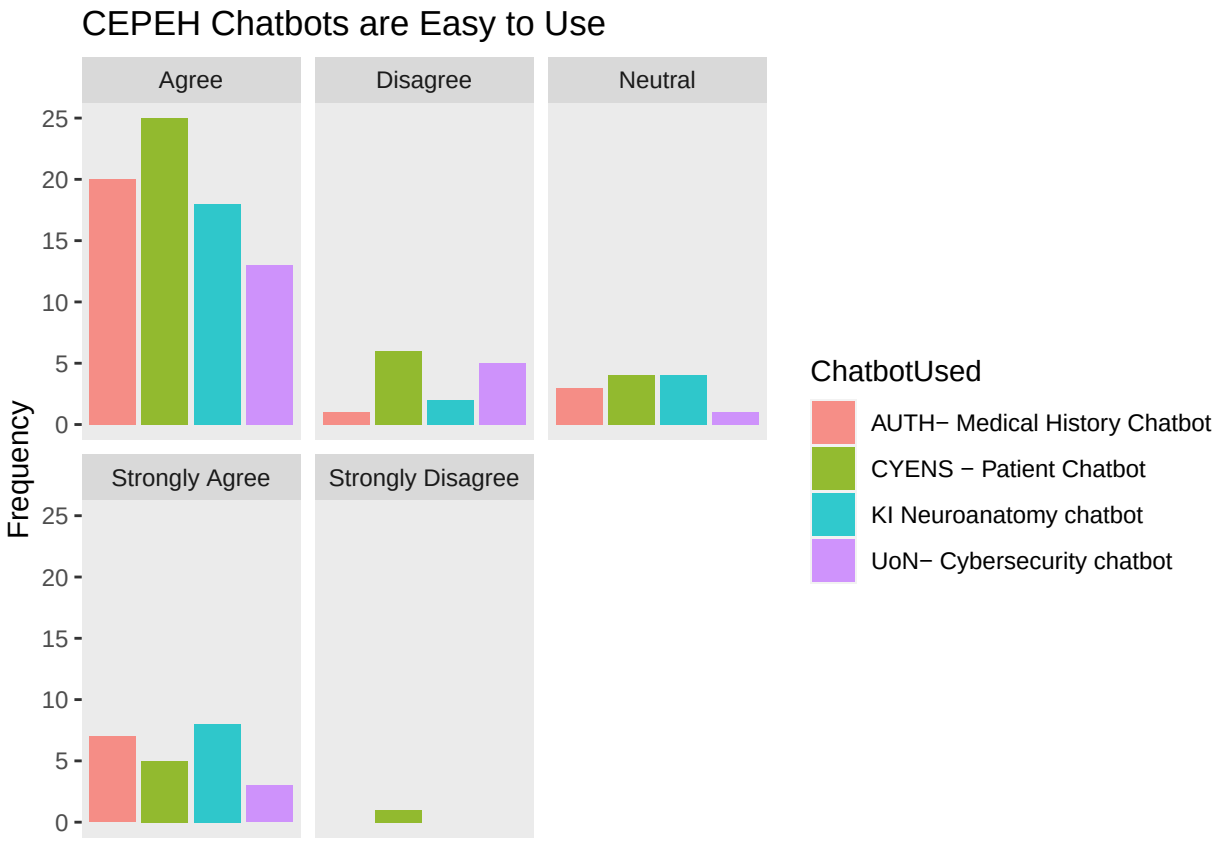


Figure 1.14: Easy to Use- Post