Applied Mathematics and Computational Intelligence
Volume 14, No. 4, 2025 [109 - 132]

Evaluation of an Ai-Powered Chatbot for University Student Support: A Framework-Based Analysis

Che Wan Shamsul Bahri Che Wan Ahmad[1*], Syed Arbaz Ahmed[2], Khirulnizam Abd Rahman[3], Syarbaini Ahmad[4], Mokmin Basri[5], Syahrul Nizam Junaini[6], Mohd Shahrul Nizam Mohd Danuri[7]

[1,2,3,4,5]Universiti Islam Selangor, 43000, Kajang, Selangor, Malaysia
[6]Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[7]Universiti Malaya, 50603 Kuala Lumpur, Malaysia

*Corresponding author: cwshamsul@uis.edu.my

Received: 2 December 2024
Revised: 27 March 2025
Accepted: 8 July 2025

ABSTRACT

This study provides a comprehensive evaluation of an AI-powered chatbot designed to assist university students with their inquiries. Using the structured and rigorous framework developed by Følstad and Taylor, the chatbot's performance is assessed based on feedback from both IT and non-IT participants. The evaluation focuses on four key dimensions: response relevance, response understandability, dialogue outcomes and dialogue efficiency. Participants responded to ten questions—nine rated on a scale of 1 to 5 and one open-ended—to provide in-depth insights into the chatbot's effectiveness, ease of use, and potential to enhance traditional student support systems. The findings highlight the chatbot's strengths, such as improved response accuracy and usability, while also identifying areas that require further refinement. This study contributes to existing research by not only assessing chatbot performance through a structured framework but also comparing the experiences of IT and non-IT users, an aspect often overlooked in prior studies. Additionally, it integrates both qualitative and quantitative evaluations to provide a well-rounded understanding of user satisfaction. The insights gained offer practical recommendations for developers and decision-makers to optimize AI-powered chatbot solutions, ultimately improving student support services and fostering a more efficient and responsive university environment.

Keywords: Artificial Intelligence (AI), Evaluation. Chatbots, Universities.

1 INTRODUCTION

Effective communication is a crucial aspect for universities, colleges or institutes aiming to attract and enrol prospective students [1]. In today's digitally driven world, where individuals are technologically proficient and accustomed to accessing information instantly through various devices, universities must meet these expectations by providing immediate responses to student queries [2]. Traditional student support channels, however, often fall short due to limitations such as slow response times, restricted operating hours, and difficulties handling high volumes of repetitive inquiries. These constraints can delay problem resolution and adversely affect the overall student experience. To address these issues, an AI-powered student inquiry chatbot was developed and

implemented using the web-based Botpress platform on the UIS website. This chatbot enables live, immediate responses to student queries, offering a more efficient alternative to conventional support channels. By leveraging this system, students can access information swiftly, thereby enhancing their overall experience. Moreover, the AI-powered chatbot is equipped with natural language processing capabilities to accurately understand and interpret student queries [3]. The development and deployment of the chatbot represent a decisive step toward improving student support, though it remains necessary to explore its efficacy and impact on user satisfaction in greater detail. This study will assess the chatbot using a qualitative analysis framework to provide deeper insights into its performance and potential benefits.

2 OBJECTIVES

This study aims to analyze chatbot interactions qualitatively to gain insights into user experiences. The framework evaluates key aspects such as performance and efficiency, serving as a foundation for future qualitative and quantitative research. It provides an initial assessment of chatbot-user conversations to support further in-depth evaluations.

3 METHODOLOGY

The messages that the chatbot and the user exchange, the metadata associated with these messages, and details on the user's engagement with interactive components in the chat dialogue serve as the data source for the analysis of the AI-powered chatbot conversation with the user [4]. We refer to a dialogue as a conversation between a user and the chatbot. The evaluation leverages Følstad and Taylor's [4] framework, focusing on key elements such as response relevance, understandability, and conversational efficiency to systematically assess the AI chatbot's performance (Table 1).

Table 1: Main framework elements [4]

Main framework Elements	Categories
Response Relevance	Relevant response
	False positive
	False negative
	Out of Scope
Response Understandability	Likely understandable
	Understandability issue
Dialogue Outcome	Relevant help - Likely used
	Relevant help – Likely not used
	Escalation offered
	No relevant help
Dialogue Efficiency	Coherent dialogue flow
	Breaks in dialogue flow

Følstad and Taylor's framework was chosen because it offers a structured and systematic way to evaluate chatbot interactions, focusing on key aspects such as response relevance, understandability, conversation outcomes, and efficiency. This approach ensures a thorough assessment of the chatbot's ability to effectively address student inquiries, aligning closely with the study's objectives.

By incorporating both qualitative and quantitative analysis, the framework provides a balanced evaluation that goes beyond measuring technical accuracy to also consider the overall user experience. Additionally, it serves as a strong foundation for identifying areas of improvement, making it highly relevant to the study's goal of enhancing AI-powered student support systems.

A design science research approach guided the development of the framework, ensuring a practical and iterative evaluation of chatbot interactions, allowing for in-depth assessment of conversational dynamics [5]. This is a suitable approach for research that seeks to design and test systems that aid problem-solving in the actual world. In this case, our framework for evaluation of an AI-powered chatbot for university student support was developed to test the system.

The framework was created in response to a practical requirement that allowed for systematic evaluation of conversation data from chatbots used for student support in order to obtain insight into the user experience [4]. To offer course for the development and assessment of the framework, the list of requirements was provided. The requirements are shown below in Table 2.

Later, the evaluation process was done which involved two categories of participants: those with an IT background and those without. There were 10 participants in each category, making up a total of 20 people. Participants of both categories used the chatbot and provided their feedback based on their experience. To this end, Google Form online survey was created, designed to measure various dimensions of interaction with the chatbot. It explored response relevance, response understandability, conversation outcomes, and conversation efficiency. Based on the literature review, Table 2 shows list of requirements of elements in a chatbot suggested by different researchers for chatbot evaluation.

Table 2: List of requirements of elements suggested by different researchers for chatbots evaluation

Elements	**Description**
Navigation & Interaction [6]	The chatbot should provide intuitive navigation and interaction for users [6].
Accuracy [7]	The classifier should correctly interpret and categorize user inputs with high accuracy [7].
Responsiveness [4][8]	The chatbot should respond promptly to user inputs, ensuring a quick and efficient interaction [4][8].
Comprehensibility [4][9]	The generated responses should be clear and easy to understand [4][9].
Realism [10]	Responses should feel natural and realistic, resembling human conversation [10].

Repetitiveness [4][11]	The chatbot should avoid repeating the same responses, offering varied and relevant answers [4][11].
Chatbot Understanding [4][12]	The chatbot should accurately understand and process user queries [4][12].
Word Error Rate [4][13]	Measure the accuracy of the chatbot in interpreting individual words within user inputs [4][13].
Concept Error Rate [8]	Measure the chatbot's accuracy in understanding the overall concepts or intentions of user inputs [8].
Appearance of Agent [14]	The visual design of the chatbot should be appealing and professional [14].
Background Colour [10][14]	The background colour should be visually pleasing and not distracting from the content [10][14].
Content Presentation [9][15]	Information should be presented in a clear and organized manner, enhancing readability and engagement [9][15].

3.1 Framework based analysis

In this section of our paper, we will explain how the evaluation of the chatbot's performance is done using the main framework elements: Response Relevance, Response Understandability, Dialogue Outcome and Dialogue Efficiency [4] as shown in Table 3 - 7. In the upcoming section, a set of questions has been prepared based on the elements in the Evaluation and Quality Assessment Section.

3.1.1 Response relevance

Table 3: Example: Response relevant

Category	Description	Example
Relevant response	The response is relevant for the user question.	User: Can I bring vehicle to campus? *Chatbot: No. Only final year students are allowed to bring vehicle.*
False positive	The response is relevant but incorrect information.	User: Can I bring vehicle to campus? *Chatbot: Yes. All students are allowed to bring vehicle.*
False negative	The response is relevant but incorrect information.	User: Can I bring vehicle to campus? *Chatbot: No. No one is allowed to bring vehicle.*

Out of scope	The response correctly indicates the user question to be out of scope	User: What is your favourite food? *Chatbot: Sorry, I was not trained to answer this question, please try again, I will try to not let you down again.*

3.1.2 Response Understandability

Table 4: Example: Response understandability

Category	Description	Example
Likely understandable	Does not indicate any understandability issue	User: How can I check my fees for current semester? *Chatbot: Please click here to check your current fees [Button with link]* User: [Clicks the button with link]
Understandability issue	User fails to use interactive elements, indicate understandability issues	User: How can I check my fees for current semester? *Chatbot: Please click here to check your current fees [Button with link]* User: How can I check fees?

3.1.3 Dialogue outcome

Table 5: Example: Dialogue Outcome

Category	Description	Example
Relevant help - Likely used	The chatbot provides assistance in line with user's objective, either through interactive features the user uses or directly in the chat.	User: How can I check my fees for current semester? *Chatbot: Please click here to check your current fees [Button with link]* User: [Clicks the button with link]
Relevant help – Likely not used	The chatbot provides assistance in line with user's objective, but user chose not to use or apply the assistance	User: How can I check my fees for current semester? *Chatbot: Please click here to check your current fees [Button with link]* [Conversation ends]
Alternative offered	The chatbot provides alternative channels of support, if there is no answer to user's question in knowledge base.	User: What is my result for this semester? *Chatbot: Sorry, I was not trained to answer this question, please try again, I will try to not let you down again. You can also try checking your result in by clicking here. [Button with link]*
No relevant help	No further assistance provided by the chatbot	User: What is my result for this semester? *Chatbot: Sorry, I was not trained to answer this question, please try again, I will try to not let you down again.* [Conversation ends]

3.1.4 Dialogue Efficiency

Table 6: Example Dialogue efficiency

Category	Description	Example
Coherent conversation flow	There are no pauses in the conversations caused by misunderstandings or inability to move the user closer to their objective.	User: How can I check my fees for current semester? *Chatbot: Please click here to check your current fees [Button with link]* User: [Clicks the button with link]
Breaks in conversation flow	There are pauses in the conversations caused by misunderstandings or inability to move the user closer to their objective.	User: How can I check my fees for current semester? *Chatbot: Please click here to check your current fees [Button with link]* User: How can I check fees?

3.2 Research questions

The aim of this study is to conduct a qualitative evaluation of AI-powered chatbots for student queries with a specific framework. Research objectives were translated into specific research questions as shown below:

I. How effectively does the AI-powered chatbot understand and respond to student inquiries on the platform? [RQ1]
II. What are the perceptions of students regarding the usability and user experience of the chatbot? [RQ2]
III. How does the performance of the AI chatbot developed compare to traditional student support channels in terms of response time and accuracy? [RQ3]

3.3 Evaluation and Quality Assessment

There were ten questions for evaluation and quality assessment (QA) asked in the online survey. These questions were constructed based on the main framework elements. The first nine questions were preset and followed a scoring system as shown in Table 7, where participants rated their experiences on a scale from 1 to 5. The evaluation questions systematically assessed key performance metrics such as response relevance, response understandability, dialogue outcomes and dialogue efficiency, providing a balanced view of user satisfaction. The final question was open-ended, inviting

participants to provide their opinions or suggestions regarding the chatbot's overall performance. The questions are listed below:

Response Relevance

- QA1: How accurately did the chatbot understand your questions?
- QA2: How often did the chatbot repeat responses?

Response Understandability

- QA3: Were the responses easy to understand?
- QA4: Responses seem natural and realistic?

Dialogue Outcome

- QA5: How satisfied are you with your experience using the chatbot?
- QA6: Rate the responsiveness of the chatbot's answer.

Dialogue Efficiency

- QA7: How easy was it to navigate and interact with the chatbot?
- QA8: Was the chatbot visually appealing and easy to use? *(Merged visual appeal with usability)*
- QA9: Was the content presented in a clear and organized manner? *(Modified for better alignment with efficiency*

Overall performance Feedback

- QA10: Please provide any suggestions or comments to improve the chatbot's performance.

3.4 Scoring Procedures

The questions (QA1 to QA9) answered by the participants in the online survey was scored using the scoring procedures as shown below in Table 7.

Table 7: Scoring Procedures.

QA	Scoring Procedures
QA1	1 – Strongly disagree the chatbot accurately understood my questions.
	2 – Disagree the chatbot accurately understood my questions.
	3 – Partially agree the chatbot accurately understood my questions.
	4 – Agree the chatbot accurately understood my questions.
	5 – Strongly agree the chatbot accurately understood my questions.

QA2	1 – Very frequently, almost every response was repeated.
	2 – Frequently, responses were repeated multiple times.
	3 – Occasionally, some responses were repeated.
	4 – Rarely, only a few responses were repeated.
	5 – Never, the chatbot did not repeat responses.
QA3	1 – Strongly disagree the responses were easy to understand.
	2 – Disagree the responses were easy to understand.
	3 – Partially agree the responses were easy to understand.
	4 – Agree the responses were easy to understand.
	5 – Strongly agree the responses were easy to understand.
QA4	1 – Strongly disagree the responses seemed natural and realistic.
	2 – Disagree the responses seemed natural and realistic.
	3 – Partially agree the responses seemed natural and realistic.
	4 – Agree the responses seemed natural and realistic.
	5 – Strongly agree the responses seemed natural and realistic.
QA5	1 – Very dissatisfied with my experience using the chatbot.
	2 – Dissatisfied with my experience using the chatbot.
	3 – Neutral, neither satisfied nor dissatisfied.
	4 – Satisfied with my experience using the chatbot.
	5 – Very satisfied with my experience using the chatbot.
QA6	1 – Very slow and unresponsive.
	2 – Slow with noticeable delays.
	3 – Average response speed.
	4 – Fast and responsive.
	5 – Very fast and highly responsive.
QA7	1 – Very difficult to navigate and interact with the chatbot.
	2 – Difficult to navigate and interact with the chatbot.
	3 – Neutral, neither easy nor difficult.
	4 – Easy to navigate and interact with the chatbot.
	5 – Very easy to navigate and interact with the chatbot.
QA8	1 – Very unattractive and poorly designed.
	2 – Unattractive and cluttered.
	3 – Neutral, neither appealing nor unappealing.
	4 – Visually appealing and well-designed.
	5 – Very visually appealing and well-structured.

QA9 1 – Strongly disagree the content was aesthetically pleasing.
2 – Disagree the content was aesthetically pleasing.
3 – Neutral, neither appealing nor unappealing.
4 – Agree the content was aesthetically pleasing.
5 – Strongly agree the content was aesthetically pleasing.

3.5 Relationship between Research Questions and Quality Assessment Questions

Table 8 and Figure 1 below shows how quality assessment (QA) questions can be mapped to the corresponding research questions (RQ):

Table 8: Mapping of research questions and quality assessment questions

Research Questions (RQs)	Quality Assessment (QA) Questions
RQ1: How effectively does the AI-powered chatbot understand and respond to student inquiries?	**QA1:** How accurately did the chatbot understand your questions? **QA2:** How often did the chatbot repeat responses? **QA3:** Were the responses easy to understand? **QA4:** Did the responses seem natural and realistic?
RQ2: What are the perceptions of students regarding the usability and user experience of the chatbot?	**QA5:** How satisfied are you with your experience using the chatbot? **QA6:** Rate the responsiveness of the chatbot's answer. **QA7:** How easy was it to navigate and interact with the chatbot? **QA8:** Was the chatbot visually appealing and easy to use? **QA9:** Was the content presented in a clear and organized manner?
RQ3: How does the chatbot compare to traditional student support channels in response time and accuracy?	**QA1:** How accurately did the chatbot understand your questions? **QA7:** How easy was it to navigate and interact with the chatbot?

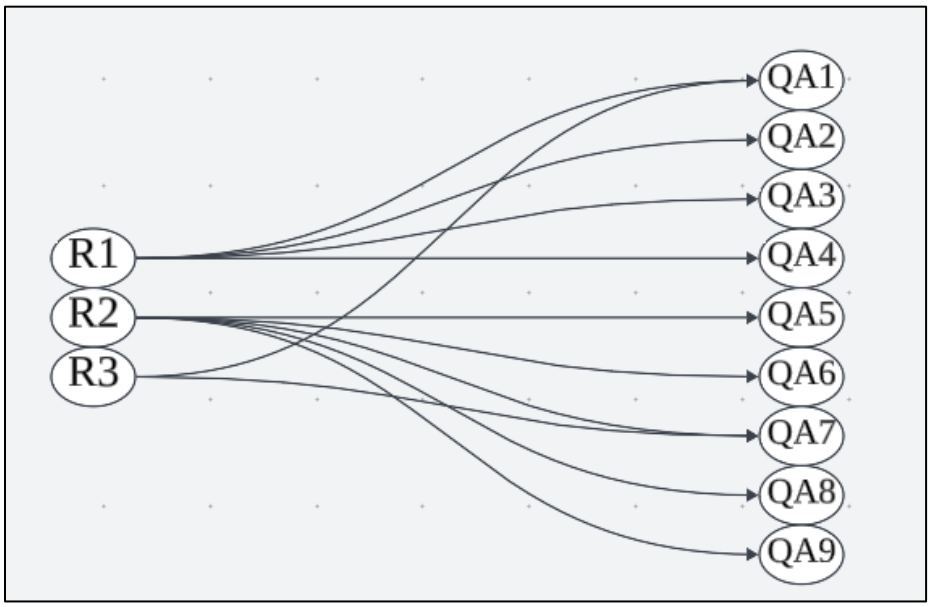

Figure 1: Mapping Between Research Questions (RQ) and Quality Assessment (QA)

3.6 Average Calculation

The formula for calculating the average score is shown (Equation 1) below:

$$\text{Average Score} = \frac{\sum_{i=1}^{n} score{=}i}{n} \quad (1)$$

Where:

- $\sum_{i=1}^{n} score = i$ is the sum of scores given by all participants for a particular question.
- n is the total number of participants which is 10 for both categories.

The formula for calculating the average for each category is shown (Equation 2) below:

$$\text{Average (category)} = \frac{\sum_{x=1}^{n} score{=}x}{y} \quad (2)$$

where:

- $\sum_{x=1}^{n} score = x$ is the sum of average score given by all participants for a particular category.
- y is the total number of questions which is 9 for both categories.

4 EVALUATION OF AI-POWERED CHATBOT RESULT

The qualitative evaluation and assessment questions that were previously constructed based on the main framework elements were conducted with two categories of people, information technology (IT) background and non-IT background from each category 10 people were chosen who interacted with the chatbot and provided feedback based on their experience. The overview was recorded in Table 9 and Table 10**.** Meanwhile, in Table 11 and 12, the feedback from participants in each category (IT and non-IT backgrounds) is recorded. Notably, 5 out of the 10 participants from each category provided feedback. Each person had the opportunity to navigate through the chatbot, ask questions and assess its performance across various criteria. Their feedback provided valuable insights into the usability, performance and overall satisfaction with the chatbot. The aggregated ratings and comments from these two categories of people offer a comprehensive overview of the chatbot's strengths and areas for improvement, guiding further iterations and enhancements to optimize the user experience.

First category of people who are from IT background are selected among the semester 2, Selangor Islamic University (UIS) students from Faculty of Creative Multimedia and Computing (FMKK). While the second category of people who are from non-IT backgrounds are randomly chosen. These people represented a diverse range of backgrounds and perspectives, ensuring that the feedback collected was comprehensive and reflective of potential user experiences.

Table 9: Average Scores for IT Background Participants

QA	Questions	Average Score (IT Background)
QA1	How accurately did the chatbot understand your questions?	3.8
QA2	How often did the chatbot repeat responses?	3.8
QA3	Were the responses easy to understand?	4.5
QA4	Did the responses seem natural and realistic?	4.1
QA5	How satisfied are you with your experience using the chatbot?	3.8
QA6	Rate the responsiveness of the chatbot's answer.	4.0
QA7	How easy was it to navigate and interact with the chatbot?	3.5
QA8	Was the chatbot visually appealing and easy to use?	4.3
QA9	Was the content presented in a clear and organized manner?	3.9
	Average	3.96

Table 10: Average Scores for Non-IT Background Participants

QA	Questions	Average Score (Non-IT Background)
QA1	How accurately did the chatbot understand your questions?	4.0
QA2	How often did the chatbot repeat responses?	3.9
QA3	Were the responses easy to understand?	4.5
QA4	Did the responses seem natural and realistic?	4.0
QA5	How satisfied are you with your experience using the chatbot?	4.0
QA6	Rate the responsiveness of the chatbot's answer.	3.7
QA7	How easy was it to navigate and interact with the chatbot?	3.3
QA8	Was the chatbot visually appealing and easy to use?	4.5
QA9	Was the content presented in a clear and organized manner?	4.0
	Average	3.98

The data is shown in Figure 2 below:

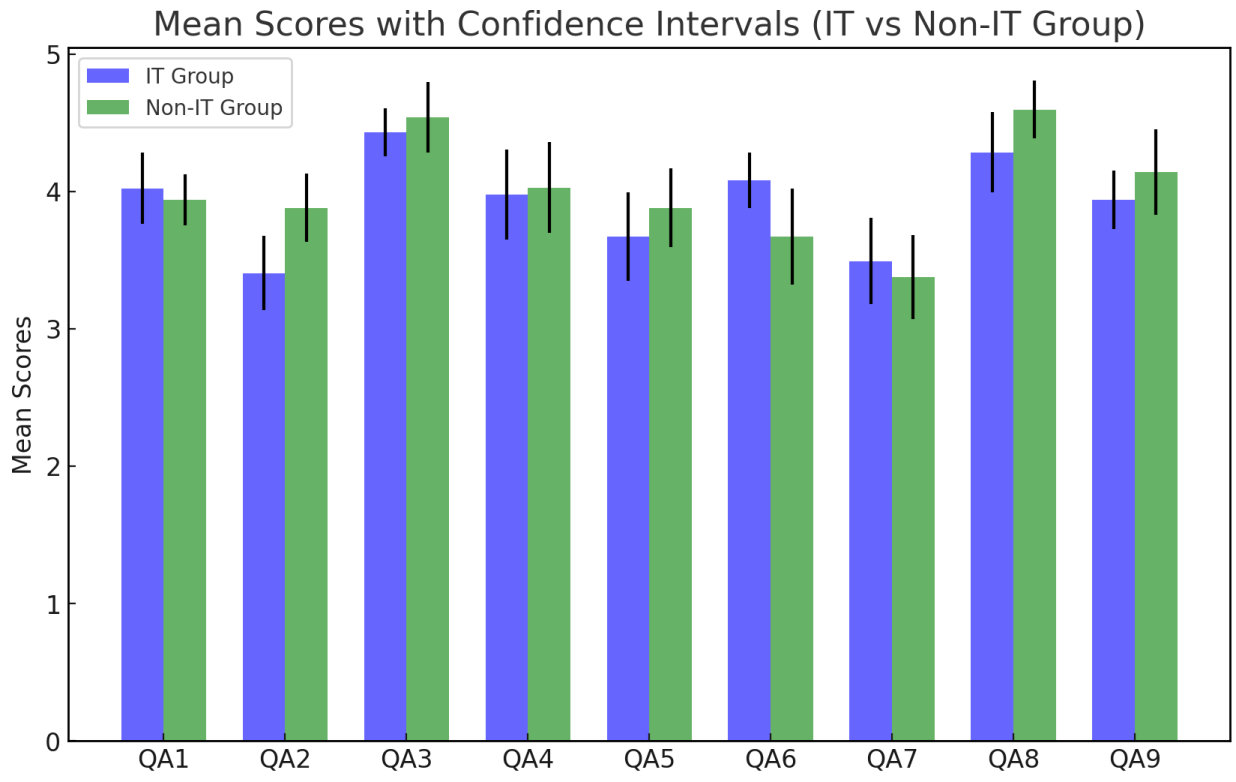

Figure 2: Means Score with Confidence Intervals (IT vs Non-IT Group)

Table 11 and Table 12 show the feedback from both IT and non-IT participants that offer valuable insights for improving the chatbot. IT users focused on technical aspects like adding a historical feature, ensuring consistent language use, and minimizing repetitive answers, while non-IT users highlighted ease of use and suggested refining the chatbot's response variety. Both groups appreciated the overall usefulness of the chatbot, with suggestions for a more attractive interface and a smoother interaction experience. This feedback shows that while the chatbot performs well, there's still room for enhancement to make it more user-friendly and efficient for all users.

Table 11: Feedback and suggestions from IT background participants

Name	Feedback/Suggestions
Ikhwan	"Maybe after this you can put a history option to make it easier for students to refer back to questions that have been asked in this chatbot"
Amirul	"Mixed language even though only asked in one language, could be improved in future."
Shamsul	"Good! can be enhanced for the next time"
Adam	"A solid chatbot that needs a little more upgrading with repetitiveness"
Ahmad	"Beautify the interface between the user and the chatbot to make it more attractive."

Table 12: Feedback and suggestions from non-IT background participants

Name	Feedback/Suggestions
Muzayyin	"A very useful chatbot."
Aqil	"Quite impressive"
Sarimah	"Overall, very good, but sometimes the chatbot asked user to provide name and telephone number again."
Ubaidillah	"I would like the chatbot to improve its variety of answers and not reject a question."

5 DISCUSSION

This research concentrates on the differences between IT and non-IT backgrounds because technical expertise can substantially influence user interactions and perceptions of an AI-powered chatbot. Users with an IT background generally possess a deeper understanding of technology, which leads them to test the chatbot with more complex, technical, or structured queries, and to hold higher expectations regarding its accuracy, efficiency, and natural language processing capabilities. In contrast, non-IT users are more likely to pose straightforward, everyday questions and prioritize ease of use over technical precision.

Furthermore, the distinct problem-solving approaches of these groups are noteworthy; IT users, being more familiar with digital systems, often navigate the chatbot using keywords or structured queries, whereas non-IT users tend to employ a more natural, conversational language, providing a broader perspective on the chatbot's ability to handle varied input styles. Given that universities cater to students from diverse academic backgrounds, it is critical to optimize chatbot design to meet the needs of all users: enhancing the underlying logic and knowledge retrieval mechanisms for IT users who might encounter technical errors and refining the user interface and response clarity for non-IT users who may struggle with usability.

Table 9 and Table 10 show the average scores obtained from participants' feedback gathered from the online survey, categorized by IT background and non-IT background, respectively. The scores explain various aspects of the chatbot's performance, in terms of its relevance and accuracy in responding to user queries and in terms of clarity and coherence of the conversation flow as well.

Finally, by addressing this often-overlooked distinction, the study fills a research gap in the chatbot literature, which typically emphasizes overall user satisfaction without considering the impact of technical background on user experience.

5.1 Relevance and Accuracy of Responses (QA1 & QA2)

This section examines the chatbot's response relevance and accuracy based on participants' average scores.

5.1.1 IT Background Participants:

The average score of the first question for QA1 is slightly below 4.0, meaning that the chatbot could have given more fitting responses to other queries. The average of the second question is the same and means that the chatbot handled out-of-scope questions to a certain extent. The scores also present the room for further improvements and the necessity for the chatbot to understand not only specific but also less common or unexpected queries.

5.1.2 Non-IT Background Participants:

As we can see from the previous table, the non-IT background group participants provided slightly more relevant answers from the chatbot with the score of 4.0 for QA1. Also, in QA2 the chatbot provided better out-of-scope answer with the score of 3.9 which is again not a perfect solution. The score differences suggest that non-IT users may have simpler expectations and communication needs, highlighting the importance of tailoring chatbot responses to varied user expertise. , which leads to better chatbot’s responses. On the other hand, IT-users might have asked more complex questions that generates unexpected chatbot’s responses.

5.2 Understandability of Responses (QA3 & QA4)

This section examines the chatbot's understandability of responses based on participants' average scores.

5.2.1 IT Background Participants:

The presence high scores in both QA3 and QA4 respectively, suggests that participants tended to find the chatbot’s responses very easy to understand. The high score of 4.5 for QA3 means the chatbot’s language and phrasing were clear and easy to understand. Closely following it is a score of 4.1 for QA4 – which means moderately easy but with slightly difference of 0.4 points. However, this indicates that there was occasional confusion or some areas where the phrasing could be better.

5.2.2 Non-IT Background Participants:

Non-IT participants also recorded high score of 4.5 for QA3, proving consistency in the chatbot's ability to communicate effectively between different user groups. A slightly lower score of 4.0 for QA4 shows more instances of confusion or unclear responses compared to the IT group, possibly due to different expectations or query complexity.

Overall, the chatbots shows consistency with this element. The small difference in QA4 between two categories scores might be due to the IT group's familiarity with more precise or technical language.

5.3 Dialogue Outcome (QA5, QA6 & QA7)

5.3.1 IT Background Participants:

They found the assistance to be quite moderately useful rating it at 3.8 for QA5 and giving a score of 4.0 for providing information (QA6). However, they did mention some shortcomings in delivering assistance giving it a score of 3.5 for QA7.

5.3.2 Non-IT Background Participants:

On the other hand, non-IT participants rated the usefulness of the chatbots information higher at 4.0 for QA5 but gave lower scores for actionable information (3.7 for QA6) and the chatbots capability to offer relevant help (3.3 for QA7).

Although they were a little forgiving when the chatbot didn't deliver support, the IT background participants thought the data that the chatbot provided was generally useful and helpful. On the other hand, people outside of the IT field said they were happy with the information's usefulness. Were less pleased by its practical guidance and the chatbot's ability to offer relevant assistance. This discovery may point to a weakness in the chatbots' ability to satisfy IT professionals' needs, who may demand more precise or contextually relevant instructions.

5.4 Dialogue Efficiency (QA8 & QA9)

5.4.1 IT Background:

The IT participants rated the flow of conversation with the chatbot at 4.3 (QA8) and the occurrence of breaks or disruptions at 3.9 (QA9), indicating a generally smooth interaction with occasional disruptions.

5.4.2 Non-IT Background:

The non-IT group gave a higher score of 4.5 for the conversation flow (QA8) and a score of 4.0 for the frequency of disruptions (QA9).

When compared to IT participants, the higher ratings provided by IT background participants for both QA8 and QA9 suggest that their interactions with the chatbot were smoother and more uninterrupted. This shows that the structure and flow of the chatbot's conversations are more

structured in term of preferences and communication patterns of IT users, who may prefer direct and ongoing communication.

5.5 Statistical Analysis

5.5.1 Descriptive Statistics

The descriptive statistics provide an overview of the central tendency and variability of scores from both IT and Non-IT participants. The mean score for IT-background participants is 3.96 (Equation 3) with a standard deviation of 0.30 (Equation 4), while the mean score for Non-IT participants is 3.98 with a standard deviation of 0.37. These values indicate that, on average, both groups rated the chatbot similarly, with only a slight difference in variation. The standard deviation values suggest that the responses were relatively consistent within each group.

The descriptive statistics summarize the central tendency and variability of scores from both IT and Non-IT participants. The **mean** $\boldsymbol{x}$ is calculated using the formula:

$$x = \frac{\sum X}{n} \tag{3}$$

where $\sum X$ is the sum of all scores, and n is the total number of participants.

For IT participants:

$$x_{IT} = \frac{3.8 + 3.8 + 4.5 + 4.1 + 3.8 + 4.0 + 3.5 + 4.3 + 3.9}{9} = 3.96$$

For Non-IT participants:

$$x_{Non-IT} = \frac{4.0 + 3.9 + 4.5 + 4.0 + 4.0 + 3.7 + 3.3 + 4.5 + 4.0}{9} = 3.98$$

The **standard deviation (SD)** is computed using:

$$\text{SD} = \sqrt{\frac{\sum(X - x)^2}{n - 1}} \tag{4}$$

For IT participants, SD_{IT} = 0.30, and for Non-IT participants, SD_{Non-IT} = 0.37.

5.5.2 Independent t-test

To determine whether the differences between the IT and Non-IT participants' scores were statistically significant, an independent t-test (Equation 5) was conducted. The resulting **t-statistic (-0.140)** and **p-value (0.890)** suggest that there is no significant difference between the two groups (Equation 6). Since the p-value is greater than **0.05**, we fail to reject the null hypothesis, meaning that any observed difference in scores could be due to random variation rather than a meaningful difference between IT and Non-IT participants. The **independent t-test** using the formula:

$$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5}$$

where:

- x_1 and x_2 are the means of IT and Non-IT groups.
- s_1^2 and s_2^2 are the variances of each group.
- n_1 and n_2 are the sample sizes.

Substituting the values:

$$t = \frac{3.96 - 3.98}{\sqrt{\frac{0.30^2}{9} + \frac{0.37^2}{9}}} \tag{6}$$

$$t = \frac{-0.02}{\sqrt{\frac{0.09}{9} + \frac{0.1369}{9}}} = \frac{-0.02}{\sqrt{0.01 + 0.0152}} = \frac{-0.02}{\sqrt{0.0252}} = \frac{-0.02}{0.159}$$

$$t = -0.140$$

The **p-value** obtained is **0.890**, which is greater than **0.05**, indicating no significant difference between IT and Non-IT participants.

5.5.3 Effect Size (Cohen's d)

The effect size, measured using **Cohen's d (-0.066)**, indicates the magnitude of the difference between the two groups (Equation 7). A Cohen's d value close to zero suggests that the effect size is very small, implying that the impact of a participant's technical background (IT vs. Non-IT) on their chatbot evaluation is negligible. This supports the t-test result, reinforcing that the differences in scores are not practically significant.

$$d = \frac{x_1 - x_2}{s_p} \tag{7}$$

where s_p is the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(9 - 1)(0.30^2) + (9 - 1)(0.37^2)}{9 + 9 - 2}}$$

$$s_p = \sqrt{\frac{8(0.09) + 8(0.1369)}{16}} = \sqrt{\frac{0.72 + 1.0952}{16}} = \sqrt{\frac{1.8152}{16}} = \sqrt{0.1134} = 0.337$$

$$d = \frac{x_1 - x_2}{s_p} = \frac{3.96 - 3.98}{0.337} = \frac{-0.02}{0.337} = -0.066$$

A **Cohen's d value of -0.066** indicates a negligible effect size.

5.5.4 Confidence Interval for Mean Difference

The confidence interval (using Equation 8) for the mean difference between IT and Non-IT participants' ratings is **(-0.36, 0.31)**. Since this range includes **zero**, it further confirms that there is no statistically significant difference between the two groups. A confidence interval that spans both positive and negative values means we cannot conclude that one group consistently rated the chatbot higher than the other. This suggests that the chatbot's usability and performance are perceived similarly across both IT and Non-IT users.

$$CI = (x_1 - x_2 \pm t_{critical} \times SE \tag{8}$$

where **SE (standard error)** is:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{9}$$

$$SE = \sqrt{\frac{0.30^2}{9} + \frac{0.37^2}{9}} = \sqrt{\frac{0.09}{9} + \frac{0.1369}{9}} = \sqrt{0.0252} = 0.159$$

Using a **t-critical value of 2.120** for a 95% confidence level:

$$CI = (x_1 - x_2 \ \pm t_{critical} \times SE \tag{8}$$

$$CI = (-0.02 \ \pm (2.120) \times 0.159$$

$$CI = (-0.02 \ \pm (0.337)$$

$$CI = (-0.36, 0.31)$$

Since **zero falls within the confidence interval**, this further confirms that there is no statistically significant difference.

5.5.5 Final Interpretation

Overall, the statistical analysis indicates that IT and Non-IT participants did not rate the chatbot significantly differently. The chatbot's usability, response quality, and overall performance were perceived consistently across both groups. This suggests that technical background does not have a strong impact on user experience, implying that the chatbot is accessible and user-friendly for individuals regardless of their familiarity with IT concepts.

6 CONCLUSION

To conclude, each research questions were answered from the quality assessment questions constructed. Below is a summary of the findings, explicitly linking the research questions (RQ) to the corresponding quality assessment (QA) criteria:

6.1 How effectively does the AI-powered chatbot understand and respond to student queries on the Botpress platform? [RQ1 – QA1, QA2, QA3, QA4]

While the AI-powered chatbot is successful at understanding and answering questions from students, as evaluated through **QA1 (accuracy of responses)** and **QA3 (clarity of responses)**. it could be more effective at handling out-of-scope queries and making sure that all of the responses are easily understood by the user as indicated by **QA2 (repetition of responses)**. It succeeds at providing clear and useful answers when it comes to questions that fall within its stated scope. While responses were generally accurate and clear, **QA4 (naturalness of responses)** suggests that further refinements could make interactions feel more intuitive and human-like.

6.2 What are the perceptions of students regarding the usability and user experience of the chatbot? [RQ2 – QA5, QA6, QA7, QA8, QA9]

Students think the chatbot is an in general helpful tool that provides a satisfying user experience as reflected in **QA5 (user satisfaction)**. However, perceptual differences, particularly between IT and non-IT users, indicate that usability needs to be improved. While QA7 (easy of navigation and interaction) pointed out several areas for improved user-friendliness, QA6 (responsiveness) verified that response speed was sufficient. Furthermore, according to QA8 (visual appeal) and QA9 (information clarity and organization), the user experience might be further enhanced by an interface that is more aesthetically pleasing and well-structured.

6.3 How does the performance of the AI chatbot developed compare to traditional student support channels in terms of response time and accuracy? [RQ3 – QA1, QA7]

In terms of **QA1 (accuracy of responses)**, the chatbot demonstrates clear advantages over traditional support systems by offering fast, accessible information with minimal disruptions, although further improvements are needed in out-of-scope query handling and response personalization. **QA7 (ease of navigation and interaction)** revealed that the chatbot was deemed efficient by both IT and non-IT users; however, non-IT participants' somewhat higher scores indicate that their interaction was less complicated, potentially because of their simpler communication preferences, lower expectations, and less knowledge of complexity of the chatbot.

6.4 Contribution of Research

Our study takes a structured approach to evaluating chatbot performance by employing a framework based on Følstad and Taylor's work. Unlike many studies that focus solely on chatbot accuracy or user satisfaction, this research systematically assesses key conversational aspects, including response relevance, understandability, conversation outcomes, and efficiency. Additionally, a unique aspect of this study is the comparison between IT and non-IT users, which provides valuable insights into how technical knowledge influences user experience—an area often overlooked in previous research.

Furthermore, while many chatbot evaluations focus on industries like customer service and healthcare, this study specifically addresses university student support, filling a gap in academic

research. The evaluation process goes beyond simple numerical ratings by incorporating qualitative feedback, offering a deeper understanding of chatbot performance than studies that rely solely on Likert-scale evaluations. More importantly, this research provides practical recommendations for improving chatbot functionality, particularly in handling out-of-scope queries, refining natural language understanding, and enhancing overall usability. By taking this comprehensive approach, the study not only contributes to existing literature but also offers actionable insights for future chatbot development.

7 SUGGESTION AND FUTURE WORK

Future work will address several key areas for enhancement. Firstly, improving the chatbot's handling of out-of-scope queries will be prioritized by expanding its natural language processing capabilities and refining the training dataset to encompass a broader range of user inquiries. This involves enlarging the training corpus with manually validated linguistic information and boosting NLP performance, thereby enabling the chatbot to better interpret and respond appropriately to off-context questions. Prior research has emphasized that chatbot effectiveness depends on well-structured training data and continuous NLP advancements to minimize errors and improve contextual understanding [16, 17]. Secondly, future versions will focus on enhancing the actionability of responses; by optimizing and further training the chatbot's knowledge base, the system can deliver responses that are not only accurate but also directly aligned with user needs. A systematic review on AI-powered chatbots highlights that integrating domain-specific knowledge bases significantly improves chatbot utility in academic environments [18].

In addition, future research should include language proficiency and chatbot experience as independent variables to determine how these factors influence user interaction and satisfaction. Testing the chatbot in multiple languages is also recommended to assess its accessibility for non-native speakers, while comparing responses between first-time and experienced chatbot users will help identify any usability gaps. Studies suggest that user background, including prior exposure to chatbot systems, plays a role in perceived usability and satisfaction [19].

Moreover, conducting longitudinal user studies with a larger and more diverse sample will provide insight into how well the chatbot continues to meet user expectations over time, allowing for the identification of trends and guiding further improvements. Research on chatbot adoption trends indicates that long-term engagement and iterative updates are essential for maintaining system relevance in educational institutions [17]. Finally, incorporating a quantitative analysis of chatbot interactions by measuring response times, resolution rates and user satisfaction scores will complement the qualitative evaluation, offering a balanced perspective that can inform ongoing tuning and optimization efforts.

ACKNOWLEDGEMENT

This research was funded by a grant from Universiti Islam Selangor (UIS) (GPIK Grant 2022/ P / GPIK /GPP-01/ 032).

REFERENCES

[1] O. Petryshyna and M. Boyko, "Communicative management in present-day university," in *Proceedings - SHS Web of Conferences*, Jan. 2021, vol. 104, pp. 1-6. doi: 10.1051/shsconf/202110402012.

[2] C. C. Ho, H. L. Lee, W. K. Lo, and K. F. A. Lui, "Developing a chatbot for college student programme advisement," in *Proceedings – International Symposium on Educational Technology*, Jul. 2018, pp. 52-56. doi: 10.1109/iset.2018.00021.

[3] M. Dibitonto, K. Leszczynska, F. Tazzi, and C. M. Medaglia, "Chatbot in a campus environment: design of LISA, a virtual assistant to help students in their university life," in *Lecture notes in Computer Science*, 2018, pp. 103–116. doi: 10.1007/978-3-319-91250-9_9.

[4] A. Følstad and C. Taylor, "Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues," *Quality and User Experience*, vol. 6, no. 6, pp. 1-17, Aug. 2021. doi: 10.1007/s41233-021-00046-5.

[5] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, Mar. 2004. doi: 10.5555/2017212.2017217.

[6] A. Huddar, C. Bysani, C. Suchak, U. D. Kolekar, and K. Upadhyaya, "Dexter the College FAQ Chatbot," in *Proceedings - International Conference on Convergence to Digital World*, Feb. 2020, pp. 1-5. doi: 10.1109/iccdw45521.2020.9318648.

[7] M. Mekni, Z. Baani, and D. Sulieman, "A smart virtual assistant for students," in *Proceedings – 3rd International Conference on Applications of Intelligent Systems*, Jan. 2020, vol. 15, pp. 1-6. doi: 10.1145/3378184.3378199.

[8] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit, and T. Theeramunkong, "A survey on evaluation methods for chatbots," in *Proceedings - 7th International Conference on Information and Education Technology*, 2019, pp. 111-119. doi: 10.1145/3323771.3323824.

[9] A. Fadhil, "Domain specific design patterns: Designing for conversational user interfaces," *arXiv*, pp. 1-7, Jan. 2018. doi: 10.48550/arxiv.1802.09055.

[10] M. Mohammed and M. M. Aref, "Chatbot system architecture," EasyChair-Preprint, Mar. 2020, [Online]. Available: https://easychair.org/publications/preprint_open/cBTw.

[11] A. Følstad and P. B. Brandtzaeg, "Users' experiences with chatbots: findings from a questionnaire study," *Quality and User Experience*, vol. 5, no. 1, pp. 1-14, Apr. 2020. doi: 10.1007/s41233-020-00033-2.

[12] P. Agrawal, A. Suri, and T. Menon, "A trustworthy, responsible and interpretable system to handle chit chat in conversational bots," *arXiv*, pp. 1-7, Jan. 2018.

doi: 10.48550/arxiv.1811.07600.

[13] S. Das and E. Kumar, "Determining accuracy of chatbot by applying algorithm design and defined process," in *Proceedings - 4th International Conference on Computing Communication and Automation*, 2018, pp. 1-6. doi: 10.1109/ccaa.2018.8777715.

[14] G. R. S. Silva and E. D. Canedo, "Towards user-centric guidelines for chatbot conversational design," *International Journal of Human-Computer Interaction*, vol. 40, no. 2, pp. 98–120, Sep. 2022. doi: 10.1080/10447318.2022.2118244.

[15] A. Fadhil and G. Schiavo, "Designing for health chatbots," *arXiv*, pp. 1-20, Feb. 2019. doi:10.48550/arxiv.1902.09022.

[16] C. W. S. B. C. W. Ahmad, K. A. Rahman, S. Ahmad and M. Basri, "Kajian penggunaan chatbot dalam institusi pendidikan: A study of the use of chatbots in educational institutions," *Malaysian Journal of Information and Communication Technology*, vol. 8, no. 2, pp. 145-157, 2003. https://doi.org/10.53840/myjict8-2-103

[17] S. A. Ahmed, M. Basri, C. W. S. B. C. W. Ahmad, K. A. Rahman and S. Ahmad, "Evolusi dan trend dalam penyelidikan chatbot: Analisis bibliometrik," *in Transformasi Digital: Inovasi Dalam Perkembangan Profesional*, Penerbit UIS, Malaysia, 2024, pp. 175-198.

[18] C. W. S. B. C. W. Ahmad, S. A. Ahmed, K. A. Rahman, S. Ahmad and M. Basri, "A systematic literature review on AI-powered chatbot for universities," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 62, no. 4, pp. 112–126, 2004. https://doi.org/10.37934/araset.62.4.111126

[19] C. W. S. B. C. W. Ahmad, S. A. Ahmed, K. A. Rahman, S. Ahmad, M. Basri, S. Abdulmana, and A. Hikmaturokhman, "A bibliometric analysis for AI-powered chatbots," *Applied Mathematics and Computational Intelligence (AMCI)*, vol. 14, no. 1, pp. 133–147, 2025. https://doi.org/10.58915/amci.v14i1.1149