Intelligent Chatbot Testing: Kuki.AI

Mounica Reddy Kandi[#1], Niranjan Reddy Masapeta[#2] , Preethi Billa[#3], Rishab Reddy Karakala[#4]

[#]*Computer Engineering Department*
San Jose State University
San Jose, California

[1]mounicareddy.kandi@sjsu.edu

[2]niranjanreddy.masapeta@sjsu.edu

[3]preethi.billa@sjsu.edu

[4]rishabreddy.karakala@sjsu.edu

***Abstract*— This document describes our testing strategy for Kuki.ai, a popular chatbot. As it stands, the testing focuses on four capacities: Domain Knowledge, ChatBot Memory, and Chatbot Pattern and Q&A. The testing strategy combines conventional testing and automation testing. The results of point-by-point testing and bug analysis are overly detailed.**

***Keywords*— Chatbot testing, AI testing, Kuki.Ai**

I. INTRODUCTION

A chatbot is an artificial intelligence based application used to converse in natural language with humans. There are different types of chatbots based on mode of communication. Siri, Alexa and Bixby are few of the voice enabled chatbots used in day to day activities for numerous useful tasks. Many of the websites have a chatbot option for frequently asked questions to customer service scenarios replacing manual labor. Advantages of chatbots include instant resolution of queries, reduced human costs and errors, 24/7 support and automating manual tasks.

Kuki (short for Mitsuku) is one of the world's most popular English language chatbots developed with an estimated users of around 25 million people over many social media sites. Kuki is an engagement-oriented chatbot in contrast to task-oriented chatbot and is capable of carrying conversations on natural things such as companionship, entertainment and other non-business use cases. Kuki is created by Steve Worswick from Pandorabots AIML technology. Kuki is capable of many functions including but not limited to dialect detection, mathematical computations and natural conversations with sentiment analysis.

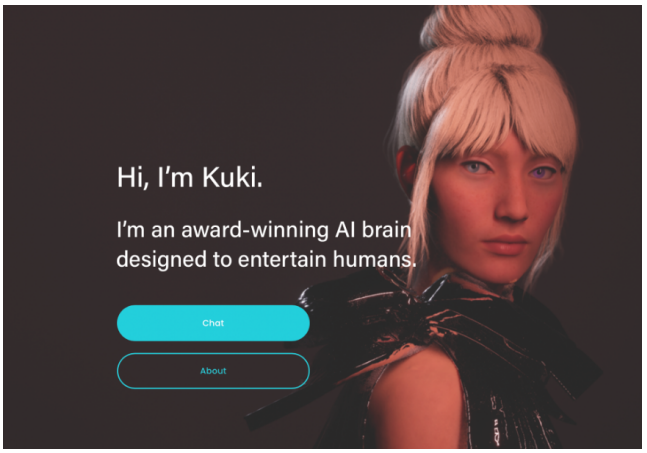

Figure 1: Kuki.AI

II. TEST INFORMATION

The scope of this project is to test the functionality of the chatbot application to evaluate the performance of Kuki. We are going to do AI testing from several aspects to test the chatbot. We have categorized the functional aspects of the chatbot into four categories: Domain Knowledge, Chatbot memory, Chatbot pattern flow, and Q&A interaction. For each category we will introduce three models namely context trees, input trees and output trees. We generate test cases based on these context trees to cover almost all scenarios. In the last section we will compare the chat output to expected output and calculate passing rate through which Kuki is evaluated.

III. TASK PARTITION

Student	**Task performed**

Mounica Reddy Kandi	Domain Knowledge testing
Preethi Billa	Chatbot memory testing
Rishab Reddy Kakakala	Chatbot pattern flow testing
Niranjan Reddy Masapeta	Q&A testing

Table 1: Task Partitioning

IV. TEST REQUIREMENTS

As shown below, functional test requirements that will be used to model AI testing will be identified and modeled. In total, there are four divisions, including the context of the text and the features considered. For each division, we identified all the possibilities and mapped them as follows. It is important that we consider all possible input combinations. These four models, which come from various points of view, can handle the majority of test case circumstances. They may concentrate on the style of input, length, punctuation, semantic meaning, and other factors. In order to compare the actual outcome with the desired ones more easily, they additionally state the expected outcome. In addition, we provide and exploit the 3-D AI-Powered Classification Decision Table. For instance on the one hand, it can view various input combinations and rules; on the other hand, it will assist us in creating a detailed test case. Below, we'll cover each of them independently.

A. Domain Knowledge

Testing the chatbot proficiency in various domains is the primary goal of domain knowledge testing. To ensure that the coverage of all relevant areas is validated, we must develop specific testing models. Furthermore, to better understand the adaptability of the product for the intended users in different scenarios, we need to consider multiple domains. The domain knowledge test allows us to determine which topics are covered, as well as to understand and assess the chatbot development direction and the most suitable target audience, and offer recommendations for further development.

B. Chatbot Memory

Chatbot memory is another critical feature used to evaluate chatbot performance. An ideal chatbot must be able to remember certain things in context or about the user to improve the efficiency of the conversation. Imagine you had an intense conversation with a chatbot regarding some issue you are facing and a couple of days later you face the same issue, it is a big hassle to feed the information again to the chatbot. Remembering the context of the conversation you had before makes your that particular day and life easy.

In this project, we will test Kuki’s chatbot memory performance which should handle the questions related to memory performance. We will test Kuki’s intelligence and ability in various memory related scenarios like long-term and short-term memory and information update scenarios.

C. Chatbot Pattern Flow

A chatbot flow is a conversational framework that plans the questions and possible answers. The chatbot frames the questions to the user based on the previous responses it got collected in the flow of the discussion. The chatbot flow is often a sequence of options the user may choose from to initiate information. For instance, the chatbot may direct the user by greeting them at the outset, or the user may take the initiative themselves. Another illustration is that the chatbot can offer matching buttons for the user to choose from when dealing with various responses and direct the chat to the following stage.

The chatbot's flexibility in responding to different kinds of interactions will also be assessed. The chatbot will be provided with a variety of conversational patterns in terms of syntax, punctuation, language, or SMS language. The output will be compared to the projected output in order to gauge the chatbot adaptability. By observing chat patterns and flow, we may comprehend the conversational flow of the chatbot.Any chatbot should, in

principle, be able to react to a wide range of diverse conversational patterns. In this setting, the chatbot ability to manage different chat patterns and participate in different interaction flows will be put to the test.

D. Q&A Interaction

Q&A (Question and Answer) interaction is one of the most basic and important functioning features of a chatbot. Q&A featured bots are super user-friendly and enhance the overall experience of chatbots. Q&A feature can be generic for chat assistants like Google Assistant, Alexa or Siri or can be domain specific according to the user requirements. In the case of domain specific chatbots, Q&A feature is helpful in answering some of the most frequently asked questions via chatbot customers may have on the website or social media page.

In our project, we are dealing with a Kuki chatbot which is an intelligent chat assistant which is supposed to act like a virtual assistant. Kuki should be able to handle most of the Q&A queries and we will test the chat application in this regard with multiple use case scenarios. The chatbot intelligence is determined by how natural the conversation is, and how well it uses its problem solving skills to respond to various queries. Queries can be of various types including but not limited to introductory questions, short questions, slang conversations and puzzle questions.

V. AI FUNCTION TEST REQUIREMENT MODELLING

As shown below, functional test requirements that will be used to model AI testing will be identified and modeled. In total, there are four divisions, including the context of the text and the features considered. For each division, we identified all the possibilities and mapped them as follows. It is important that we consider all possible input combinations. These four models, which come from various points of view, can handle the majority of test case circumstances. They may concentrate on the style of input, length, punctuation, semantic meaning, and other factors. In order to compare the actual outcome with the desired ones more easily, they additionally state the expected outcome. In addition, we provide and exploit the 3-D AI-Powered Classification Decision Table. For instance on the one hand, it can view various input combinations and rules; on the other hand, it will assist us in creating a detailed test case. Below, we'll cover each of them independently.

A. Input Context Model

The Input Context Model concentrates on identifying various input settings. As an illustration, some chatbots allow for input in the form of text, graphics, hyperlinks, and multiple languages. It can only input text for kuki.ai. There are, however, several venues for text input. For instance, using domain knowledge, we can identify several languages and their use of punctuation, capitalization, and presentational styles.

B. Input Classification Model

We are concentrating on various input content types for the input classification model. The classification can change significantly depending on the testing criteria. And it ought to concentrate on the functions. For instance, we should incorporate several emotional types while testing a chatbot capability to provide emotional support. A chatbot named Kuki.ai, which strives to behave like a real person and a friend, has features that allow it to react to questions about greetings, light conversational subjects, emotional support, and common knowledge. Therefore, we should set more specific testing items and include these types in the input classification model for domain knowledge.

C. Output Classification Model

The AI-powered model's output model is intended to establish the categorization and identify various outputs. The two categories of output that are typically required are Valid and Invalid. And for both of these types, we need more precise types. For instance, an

incorrect output could include a false answer, no response, or a response that is unrelated. We can also have various kinds of expectations for legitimate output.

VI. AI TEST MODELING FOR TEST CATEGORIES

A. Input Classification Model for testing

VI.A.1 Input Model: Domain Knowledge

Several domains can be considered to fully comprehend how well the product will adapt to the intended users in various scenarios. We have considered movies under domain knowledge to check how well Kuki.ai knows about movies. Therefore, we have considered the below four categories for kuki.ai.

- Genre
- Language
- Viewer Restriction
- Rating

The first category will be Genre, where we will test knowledge about different genres. We have multiple genres, but we are just considering a few genres like comedies, horror, romance, and thrillers. Per year thousands of movies are released across the world, and these movies can be released in different languages. Our second category is the language, where we are considering three languages for testing purposes (English, Hindi, and Korean).

Furthermore, we'll also consider the Viewer Restriction, where people below 18 years of age are not allowed to watch all the movies, whereas people above 18 can watch all the movies. The last category would be Rating. Based on the viewer reviews, the Rating can be low, medium, or high.

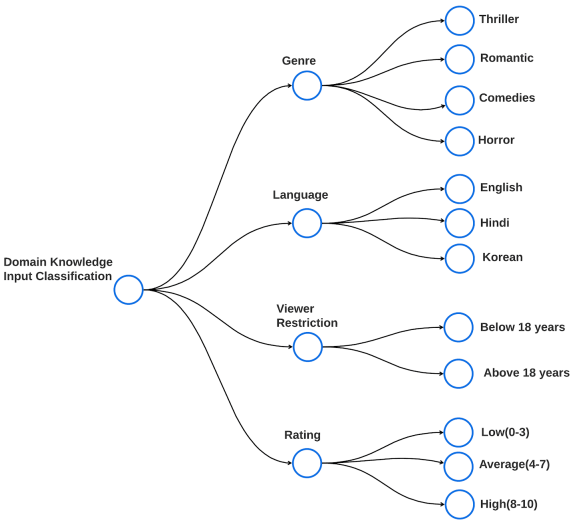

Figure 2: Input classification tree for Domain Knowledge

VI.A.2 Input Model: Chatbot Memory

You may test a chatbot memory efficiency by posing Kuki questions about various topics. We anticipate that the chatbot will be able to engage in basic one-on-one chats with users to learn about them, assess the general information they have provided, and learn about their opinions. The User information category can be further divided into a number of subcategories, including the User's personal information, Popular Likes and Dislikes, and Related Information.

In the second category, which is where we can converse and ask questions about general topics associated with various categories, such as geographic information, political knowledge, and technological-based queries, we take into account the generic information that the user has provided. The chatbot ability to recall the essential details the user provided will be put to the test in this section.

The final testing subcategory focuses on user feedback provided to Kuki. This category can be further broken down into previously provided information, travel background, and educational background. There may be other user opinions shared with Kuki, but we just selected a small number of them to test in light of the project preview.

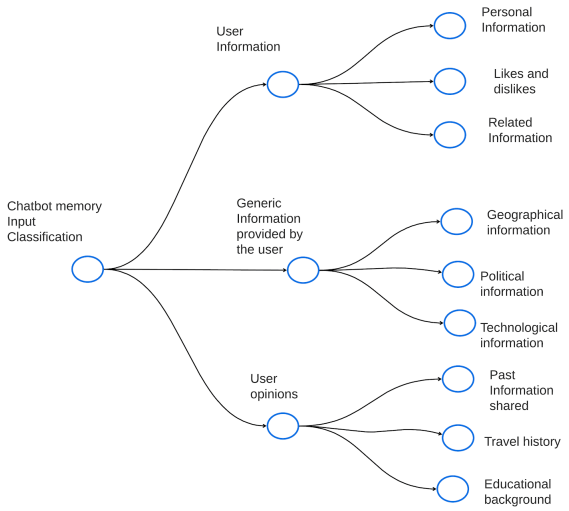

Figure 3: Input classification tree for Chatbot Memory

VI.A.3 Input Model: Chatbot Pattern

In the chatbot context model we have considered four different types of contexts. Entities are the first one where it describes whether the chatbot can recognize the keywords and the vocabulary in the chat pattern. Second is the continuity, it mainly focuses on the ability of the chatbot to switch the context. It describes the potential of the chatbot to maintain smooth communication and test whether the chatbot can quickly adapt to the new context change.

Next context is about the purpose. It checks the chatbots performance on the user specific tasks. Such as whether the user is using the chatbot for searching some information or to get any recommendations or for daily tasks. The last one is about the response time. The time consumed by the chatbot in order to respond to the user queries. It's the chatbot capability to respond in minimum time. This is the model where it entirely focuses on the chatbot context in the chart pattern and conversation flow. This model tests the chatbot's performance on various aspects while the user starts the conversation with the chatbot. Whether the chatbot is able to recognize the chat patterns in the user conversation texts.

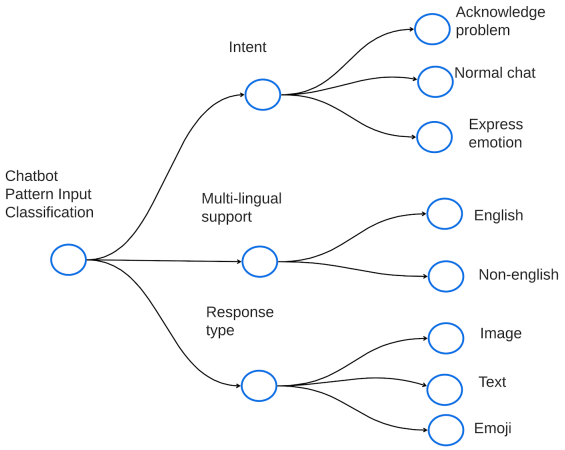

Figure 4: Input classification tree for Chatbot Pattern

VI.A.4 Input Model: Q&A Interaction

We identified three different categories of questions to test Kuki chatbot Q&A performance. We expect the chatbot to be able to have simple one-on-one questions and answers for all purposes. Chatbot should be able to handle all types of questions like what, why, where and who with all user sentiments.

We have divided the input classification into two categories based on the question type and intent of the question. Sentiment can be positive or negative or neutral depending on the user's emotion.

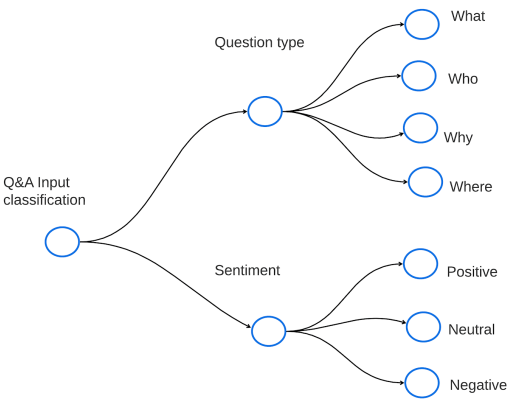

Figure 5: Input classification tree for Q&A interaction

B. Input Context Model for testing

VI.B.1 Context Model: Domain Knowledge

For domain knowledge, we've defined two context areas: character and type. We use these two context areas to determine the input category and check the chatbot response to the input. The type defines if the input given for the domain knowledge testing is either narrative or a question, i.e., to see if you can have a normal conversation with Kuki.ai or ask direct questions. We must observe if Kuki responds to only a sentence with letters or a sentence with special characters along with the letters in the input.

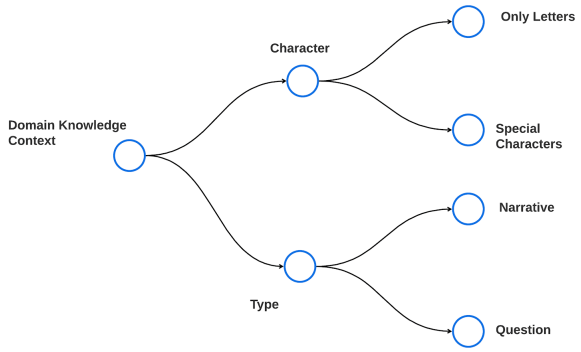

Figure 6: Context tree for Domain Knowledge

VI.B.2 Context Model: Chatbot Memory

The two types of input inquiry—information user, memory, can be used to test the input tree for chatbot memory. Information can be classified into two categories as current and updated information. The memory, which is the second category, is the main category in the input. By posing the questions that were previously discussed, it examines the kuki's short and long term memory.

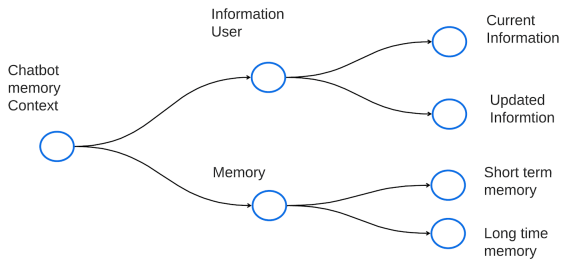

Figure 7: Context tree for Chatbot Memory

VI.B.3 Context Model: Chatbot Pattern

This model is used for testing the chatbot features efficiently. In this we consider the input given by the user to the application. Based on the input we have classified it into five different categories. The first one we would like to test is about the task specific. Which projects about the user intention in order to use the application. Which includes whether the user is using the chatbot for knowing information or to get any suggestion from the chatbot regarding a situation. The second category is about the intent of the conversation in which we perform testing on the chatbot capability when the user does a casual chat. How the chatbot responds to the user when they express their feelings and emotions. Will the chatbot acknowledge a problem?. In some situations the user is interested in knowing the solutions from the chatbot for the problems they face.

We also test for the multilingual support provided by the chatbot in the third category. When the user gives input other than english language. What kind of output the chatbot generates shows the multi language support of the chatbot.

In the fourth category we test for the response of the chatbot for different kinds of input. When the user gives the input in different formats such as text, image, idioms, etc. Does the chatbot recognize these different input formats and give the correct output. In the last category we test on the length of the input given by the user to the chatbot. Whether the input is a single word or a single sentence or the input consists of multiple sentences. In all these cases we consider the output generated by the chatbot.

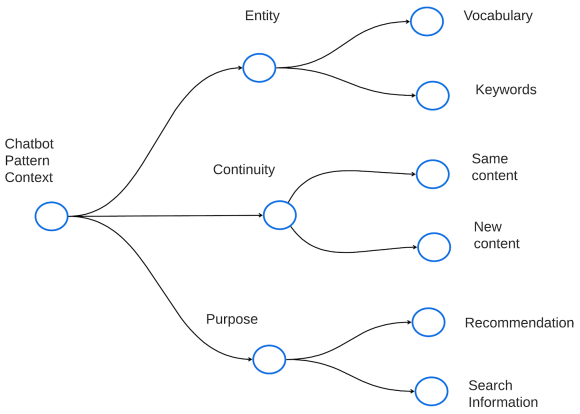

Figure 8: Context tree for Chatbot Pattern

VI.B.4 Context Model: Q&A Interaction

The context tree for Q&A interaction can be divided into two categories: Alphanumeric and Syntax type. Alphanumeric context covers all possible questions with characters including A-Z, a-z, 0-9 and special characters such as @,$,*,&,+ etc,. Syntax of the question can be broadly classified into Valid syntax, Invalid Grammar and Incorrect Spelling.

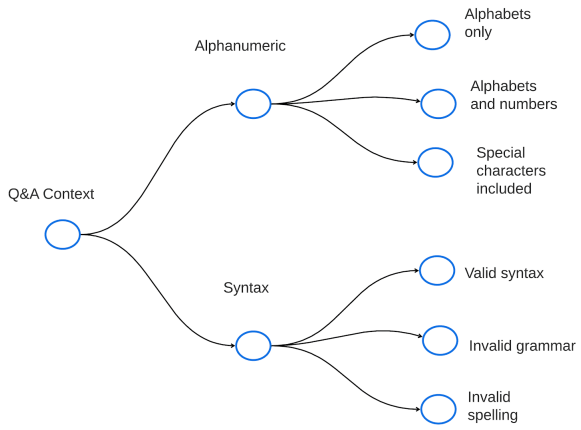

Figure 9: Context tree for Q&A Interaction

C. Output Classification Model for testing

For our project, there are different categories of output. We can decide whether the response for the given input is Relevant, Irrelevant, Correct, No Response, or an Error Message.

We can consider a result relevant if the output generated is almost relevant to the given input and Irrelevant if the output is not relevant to the input. The outcome will be Correct if we get the exact answer for the given input.

The output can be a No Response or Error message if the chatbot doesn't know what we asked.

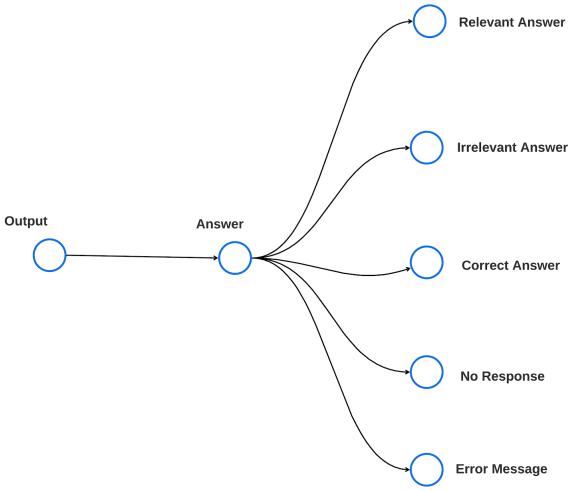

Figure 10: Output Classification Model

D. AI-Powered Classification Decision Tables

VI.D.1 Domain Knowledge

Input		R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
Genre	Thriller	T	F	F	F	F	F	F	F	F	F	F	F
	Romantic	F	T	F	F	F	F	F	F	F	F	F	F
	Comedies	F	F	T	F	F	F	F	F	F	F	F	F
	Horror	F	F	F	T	F	F	F	F	F	F	F	F
Language	English	F	F	F	F	T	F	F	F	F	F	F	F
	Hindi	F	F	F	F	F	T	F	F	F	F	F	F
	Korean	F	F	F	F	F	F	T	F	F	F	F	F
Viewer Restriction	Below 18 years	F	F	F	F	F	F	F	T	F	F	F	F
	Above 18 years	F	F	F	F	F	F	F	F	T	F	F	F
Rating	Low (0-3)	F	F	F	F	F	F	F	F	F	T	F	F
	Average (4-7)	F	F	F	F	F	F	F	F	F	F	T	F
	High (8-10)	F	F	F	F	F	F	F	F	F	F	F	T
Input Context													
Character	Only letters	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Contains special characters		✓		✓	✓		✓	✓	✓	✓		✓
Type	Narrative			✓					✓	✓			
	Question	✓	✓		✓	✓	✓	✓			✓	✓	✓
Output Context													
Answer	Irrelevant Answer		✓							✓		✓	✓
	No response				✓								
	Error Message							✓					
	Relevant Answer	✓				✓	✓				✓		
	Correct Answer			✓					✓				

Table 2: Domain Knowledge 2D decision table

Figure 11: Domain Knowledge 3D decision table

Figure 12: Chatbot Memory 3D decision table

VI.D.2 Chatbot Memory

Input		R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
User Information	Personal information	T	F	F	F	F	F	F	F	F	T	F	F
	Likes and dislikes	F	T	F	F	F	F	F	F	F	F	T	F
	Relation information	F	F	T	F	F	F	F	F	F	F	F	T
Generic Information	Geographical information	F	F	F	T	F	F	F	F	F	F	F	F
	Political information	F	F	F	F	T	F	F	F	F	F	F	F
	Technological information	F	F	F	F	F	T	F	F	F	F	F	F
User Opinions	Past information	F	F	F	F	F	F	T	F	F	F	F	F
	Travel History	F	F	F	F	F	F	F	T	F	F	F	F
	Education background	F	F	F	F	F	F	F	F	T	F	F	F
Input Context													
Information user	Current information	✅						✅			✅		
	Updated info		✅		✅				✅				
Memory	Short term memory			✅		✅				✅			
	Long term memory						✅					✅	✅
Output Context													
Answers	Irrelevant Answer	✅	✅			✅		✅					
	No Response									✅		✅	✅
	Error Message						✅						
	Relevant Answer										✅		
	Correct Answer												

Table 3: Chatbot Memory 2D decision table

VI.D.3 Chatbot Pattern

Input		R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
Intent	Acknowledge problem	T	F	F	F	F	F	F	F	F	T	F	F
	Normal Chat	F	T	F	F	F	F	F	F	F	F	T	F
	Express emotion	F	F	T	F	F	F	F	F	F	F	F	T
Multi-lingual support	English	F	F	F	T	F	F	F	F	F	F	F	F
	Non-english	F	F	F	F	T	F	F	F	F	F	F	F
Response Type	Image	F	F	F	F	F	F	T	F	F	F	F	F
	Text	F	F	F	F	F	F	F	T	F	F	F	F
	Emoji	F	F	F	F	F	F	F	F	T	F	F	F
Input Context													
Entity	Vocabulary	✅						✅			✅		
	Key words		✅		✅				✅				
Continuity	Same content			✅		✅				✅			
	New content						✅					✅	✅
Purpose	Recommendation								✅				
	Search Information										✅		✅
Output Context													
Answers	Irrelevant Answer	✅	✅			✅		✅					
	No Response									✅		✅	✅
	Error Message						✅						
	Relevant Answer										✅		
	Correct Answer												✅

Table 4: Chatbot Pattern 2D decision table

Figure 13: Chatbot Pattern 3D decision table

Figure 14: Q&A Interaction 3D decision table

VI.D.4 Q&A Interaction

Input		R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
Question Type	What	T	F	F	F	F	F	F	F	F	T	F	F
	Who	F	T	F	F	F	F	T	F	F	F	T	F
	Why	F	F	T	F	F	F	F	F	F	F	F	T
	Where	F	F	F	T	F	T	F	F	F	T	F	F
Sentiment	English	F	F	F	T	F	F	F	F	F	F	F	F
	Non-english	F	F	F	F	T	F	F	F	F	F	F	F
Input Context													
Alphanumeric	Alphabets only	✅						✅			✅		
	Alphabets and numbers		✅		✅				✅				
	Special characters included												
Syntax	Valid syntax			✅		✅				✅			
	Invalid grammar	✅					✅					✅	✅
	Invalid spelling				✅				✅				
Output Context													
Answers	Irrelevant Answer	✅	✅			✅		✅					
	No Response									✅		✅	✅
	Error Message						✅						
	Relevant Answer										✅		
	Correct Answer												✅

Table 5: Q&A Interaction 2D decision table

VII. CONVENTIONAL TEST REPORT AND ANALYSIS

A. Domain Knowledge

- Input Classification: 12
- Input Context: 4
- Total possible cases: 48
- No of used test cases: 21
- Test cases passed: 14
- Total cases failed: 7

Input Classification	No. of test cases	Passed	Passed Rate	Failed	Failed Rate
Genre	5	4	80%	1	20%
Language	7	5	71.4%	2	28.5%
Viewer Restriction	5	3	60%	2	40%
Rating	4	2	50%	2	50%
Total	**21**	**14**	**66.66%**	**7**	**33.33%**

Table 6: Conventional Test Report Domain Knowledge

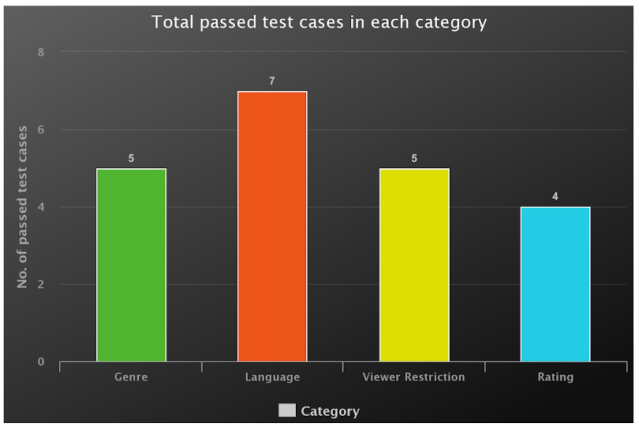

Figure 15: Passed test cases in each category for Domain Knowledge

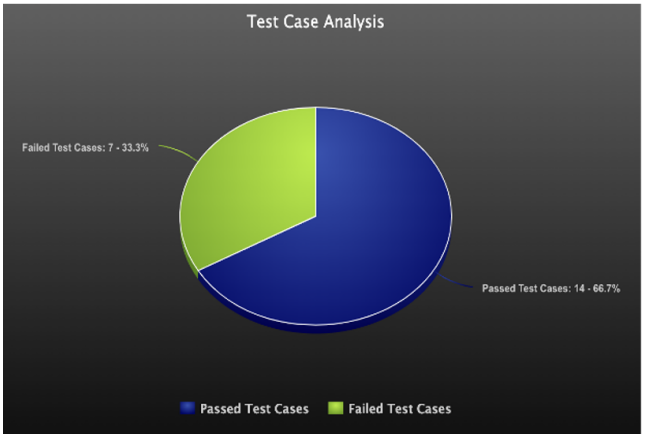

Figure 16: Test case Analysis Domain Knowledge

B. Chatbot Memory

- Input Classification: 9
- Input Context: 4
- Total possible cases: 36
- No of used test cases: 12
- Test cases passed: 8
- Total cases failed: 4

Input Classification	No. of test cases	Passed	Passed Rate	Failed	Failed Rate
User Information	6	4	75%	2	25%
Generic Information	3	2	66.66%	1	33.33%
User Opinion	3	2	66.66%	1	33.33%
Total	**12**	**8**	**75%**	**4**	**25%**

Table 7: Conventional Test Report Chatbot Memory

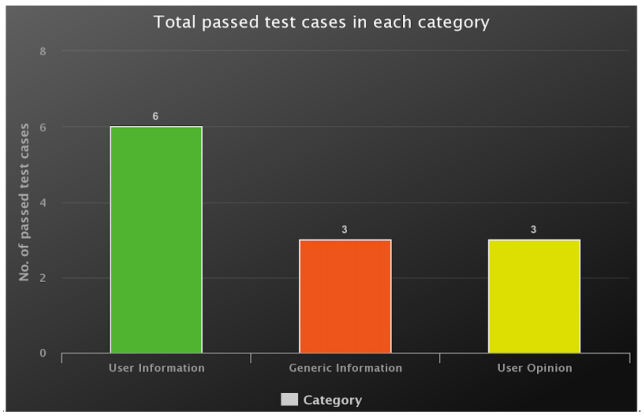

Figure 17: Passed test cases in each category for Chatbot Memory

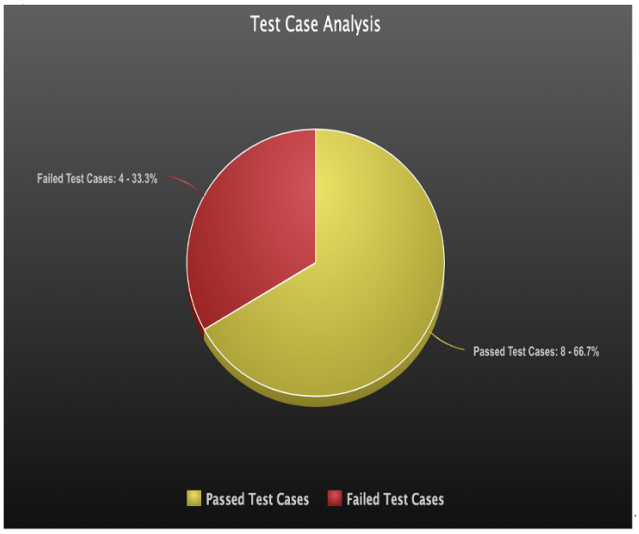

Figure 18: Test case Analysis Chatbot Memory

C. Chatbot Pattern

- Input Classification: 8
- Input Context: 6
- Total possible cases: 48
- No of used test cases: 9
- Test cases passed: 5
- Total cases failed: 4

Input Classification	No. of test cases	Passed	Passed Rate	Failed	Failed Rate
Intent of the conversation	3	2	66.6%	1	33.3%
Multilingual support	2	1	50%	1	50%
Response type	4	2	50%	2	50%
Total	**9**	**5**	**55.55%**	**4**	**44.44%**

Table 8: Conventional Test Report Chatbot Pattern

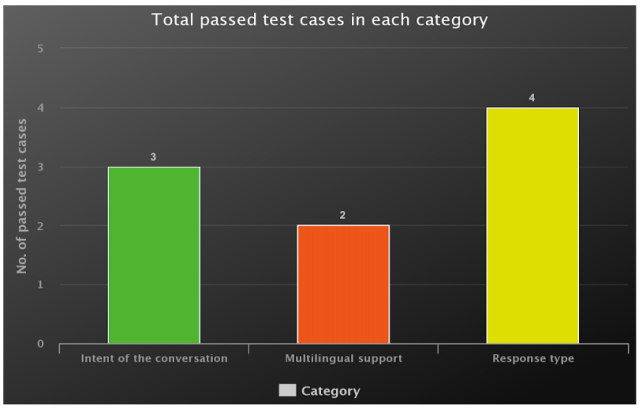

Figure 19: Passed test cases in each category for Chatbot Pattern

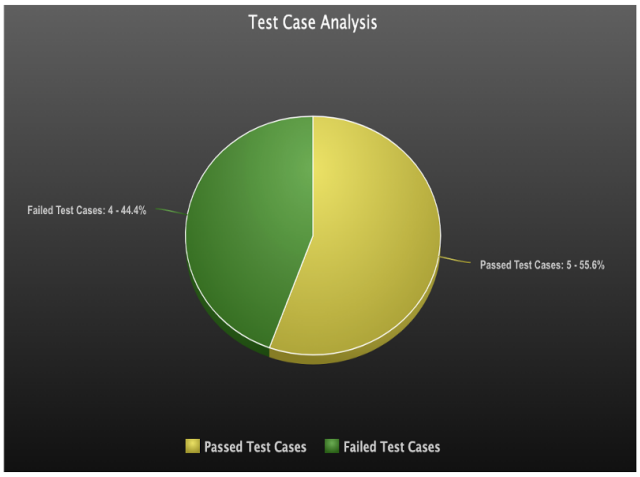

Figure 20: Test case Analysis for Chatbot Pattern

D. Q&A Interaction

- Input Classification: 7
- Input Context: 6
- Total possible cases: 42
- No of used test cases: 12
- Test cases passed: 10
- Total cases failed: 2

Input Classification	No. of test cases	Passed	Passed Rate	Failed	Failed Rate
Question Type	12	10	83.33%	2	16.66%
Sentiment					
Total	**12**	**10**	**83.33%**	**2**	**16.66%**

Table 9: Conventional Test Report Q&A Interaction

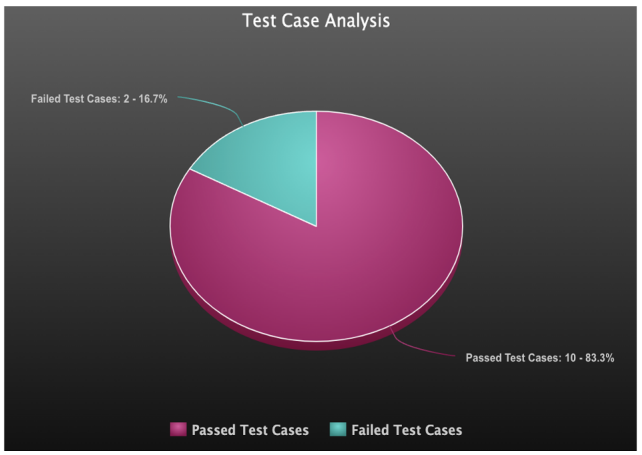

Figure 21: Test case Analysis for Q&A Interaction

VIII. AI TESTING ANALYSIS

The following is the AI test complexity for each category:

- Domain Knowledge: 216 cases
- Chatbot Memory: 108 cases
- Chatbot Pattern: 144 cases
- Q/A Interaction: 108 cases

Total Complexity: 576 cases

A. Domain Knowledge

```
Total positive results for Domain Knowledge Testing: 146
Total negative results for Domain Knowledge Testing: 70
Total time consuming for Domain Knowledge Testing: 1618
```

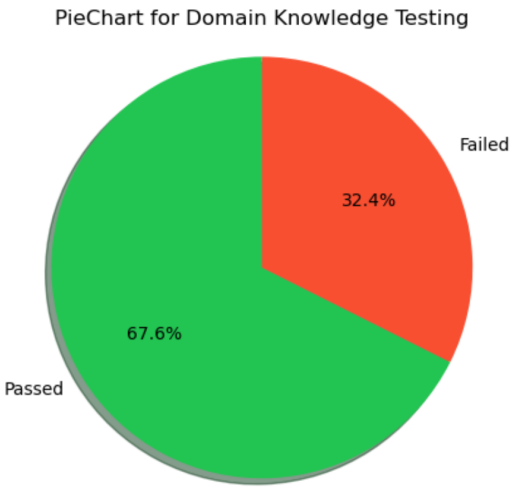

Figure 22: Automation Test Analysis for Domain Knowledge

B. Chatbot Memory

```
Total positive results for Chat Memory Testing: 57
Total negative results for Chat Memory Testing: 50
Total time consuming for Chat Memory Testing: 789
```

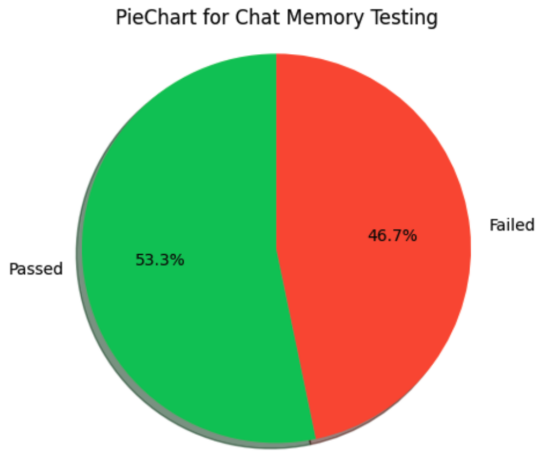

Figure 23: Automation Test Analysis for Chatbot Memory

C. Chatbot Pattern

```
Total positive results for Chat Pattern Testing: 114
Total negative results for Chat Pattern Testing: 35
Total time consuming for Chat Pattern Testing: 1124
```

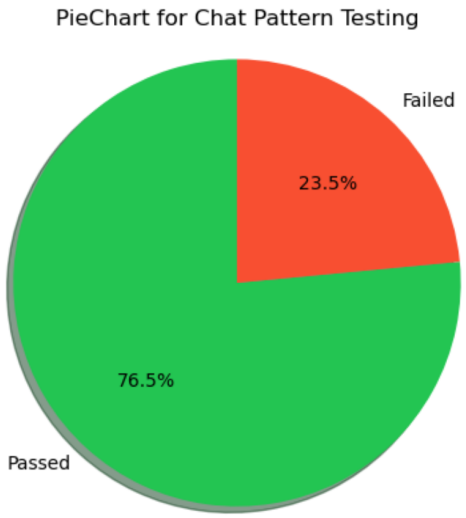

Figure 24: Automation Test Analysis for Chatbot Pattern

D. Q&A Interaction

```
Total positive results for QA Testing: 45
Total negative results for QA Testing: 63
Total time consuming for QA Testing: 831
```

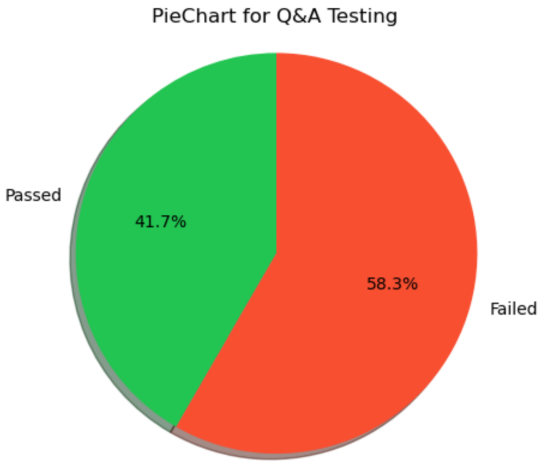

Figure 25: Automation Test Analysis for Q&A Interaction

	Testing Cases	Pass Cases	Fail Cases	Pass Rate	Time Consuming
Domain Knowledge	216	146	70	67.6%	1618 seconds
Chatbot Memory	108	58	50	53.3%	789 seconds
Chat Flow	148	114	35	77%	1124 seconds
Q&A interaction	108	45	63	41.7%	831 seconds

Table 10: Automation Test Analysis

IX. CONVENTIONAL VS AI TESTING ANALYSIS

A. Domain Knowledge

Domain Knowledge	Conventional	Automation
Total test cases	21	216
Passed test cases	14	146
Failed test cases	7	70
Total time consumed	29.7 min	26.97 min
Time per test case	1.4 min	0.11 min

Table 11: Domain Knowledge Conventional vs AI Testing Analysis

B. Chatbot Memory

Chatbot Memory	Conventional	Automation
Total test cases	12	108
Passed test cases	8	58
Failed test cases	4	50
Total time consumed	17.2 min	13.1 min
Time per test case	1.4 min	0.12 min

Table 12: Chatbot Memory Conventional vs AI Testing Analysis

C. Chatbot Pattern

Chatbot Pattern	Conventional	Automation
Total test cases	9	148
Passed test cases	5	114
Failed test cases	4	35
Total time consumed	19.4 min	18.7 min
Time per test case	2.1 min	0.13 min

Table 13: Chatbot Pattern Conventional vs AI Testing Analysis

D. Q/A Interaction

Q/A	Conventional	Automation
Interaction		
Total test cases	12	108
Passed test cases	10	45
Failed test cases	2	63
Total time consumed	15.7 min	13.85 min
Time per test case	1.3 min	0.13 min

Table 14: Q&A Interaction Conventional vs AI Testing Analysis

X. CONCLUSION

Kuki chatbot application testing has been successfully demonstrated in this project. We divided the testing categories into four categories: Domain Knowledge, Chatbot Memory, Chatbot Pattern, Q/A interaction. We designed classification models for input, context, and output trees. Test cases are designed using AI tools by generating 2D and 3D decision tables. Comparison between conventional testing and AI testing has been performed to compare and analyse both the results. Overall, Kuki performed well on domain knowledge and chatbot pattern categories. Kuki performed poorly on chatbot memory and q/a interaction as compared to normal AI voice applications like google assistant, siri or alexa. It is given as those applications are connected to an internet database to get results from.

In future works, automation scripts can be made more efficient to prevent crashes and to dynamically change the wait time based on the response given by the application.

XI. ACKNOWLEDGEMENT

We would like to thank Professor Gao for giving us an opportunity to work on this project and guide us through the same. We strongly believe we have gained immense knowledge and research experience working through the project.