



**SAN JOSÉ STATE
UNIVERSITY**

Deliverable #2(B) - AI Test Report

Chatbot AI Testing - Kuki.AI

Group 4

Mounica Reddy Kandi
Preethi Billa
Rishab Reddy Karakala
Niranjan Reddy Masapeta

Table of Contents

1. AI Function Test Requirement Analysis	4
1.1. AI function test requirement	4
1.1.1. Domain Knowledge	4
1.1.2. Chatbot Memory	4
1.1.3. Chatbot Pattern	5
1.1.4. Q&A Interaction	5
1.2. AI function test requirement modeling	6
1.2.1. Input Context Model	6
1.2.2. Input Classification Model	6
1.2.3. Output Model	6
1.2.4. AI-powered Decision Table	7
2. AI Test Modeling for Selected AI Features	7
2.1. Context modeling for each selected AI-powered function/feature	7
2.1.1. Context tree for Domain Knowledge	7
2.1.2. Context tree for Chatbot memory	8
2.1.3. Context tree for Chatbot Pattern	9
2.1.4. Context tree for Q&A interaction	11
2.2. AI-powered function input classifications	11
2.2.1. Input tree for Domain Knowledge	11
2.2.2. Input tree for Chatbot memory	14
2.2.3. Input tree for Chatbot Pattern	15
2.2.4. Input tree for Q&A interaction	17
2.3. AI-powered function output/event/action classifications	18
2.4. AI-powered function classification decision tables	18
2.4.1. Decision table for Domain Knowledge	18
2.4.2. Decision table for Chatbot memory	22
2.4.3. Decision table for Chatbot Pattern	26
2.4.4. Decision table for Q&A interaction	29
3. AI Function Test Cases with Inputs/Expected Outputs	33
3.1. Test data models	33
3.1.1. Domain Knowledge data model	33
3.1.2. Chatbot Memory data model	35
3.1.3. Chatbot Pattern data model	36
3.1.4. Q&A interaction data model	37
3.2. Test case reports	38
3.2.1. Domain Knowledge test cases	38
3.2.2. Chatbot Memory test cases	46
3.2.3. Chatbot Pattern test cases	51
3.2.4. Q&A interaction test cases	57

3.3. Test case analysis	65
3.3.1. Domain Knowledge test case analysis	65
3.3.2. Chatbot Memory test case analysis	65
3.3.3. Chatbot Pattern test case analysis	66
3.3.4. Q&A interaction test case analysis	67
3.3.5. Test case analysis summary	68
4. AI Function Test Results and Test Criteria	69
4.1. AI test model-based test results analysis test complexity, and test result statistics	69
4.1.1. Analysis for Domain Knowledge	69
4.1.2. Analysis for Chatbot memory	70
4.1.3. Analysis for Chatbot pattern flow	70
4.1.4. Analysis for Q&A interaction	71
4.1.5. Time Complexity	71
4.2. AI Function bug analysis	72
4.2.1. Domain Knowledge bug analysis	72
4.2.2. Chatbot Memory bug analysis	76
4.2.3. Chatbot pattern and flow bug analysis	78
4.2.4. Q&A Interaction bug analysis	81
4.3. AI function test quality assessment	83

1. AI Function Test Requirement Analysis

1.1. AI function test requirement

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.

1.1.1. Domain Knowledge

Testing the chatbot's 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's development direction and the most suitable target audience, and offer recommendations for further development.

1.1.2. 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, paraphrased questions, and information update scenarios. To pass this test, chatbot should ideally pass 65% of test case scenarios.

1.1.3. Chatbot Pattern

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's 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's ability to manage different chat patterns and participate in different interaction flows will be put to the test.

1.1.4. 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's 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. To pass this test, chatbot should be able to pass at least 75% of the scenarios tested.

1.2. AI function test requirement modeling

1.2.1. 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.

1.2.2. 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's 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.

1.2.3. Output 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.

1.2.4. AI-powered Decision Table

A three-dimensional cube table with three sides called the 3-D AI-Powered Classification Decision Table displays all rules/combinations based on your model (Input Context Model, Input Classification Model and Output Model). Each of your models' ids are always listed in the first column, and the subsequent columns will list each variable in turn. Since there could be up to $2N$ rules—where N is the number of the variable—and the cube surface has a limited amount of room, not every rule will be displayed in the 3-D table. However, once we have the decision table, we can produce all test cases and base our own test cases on them.

2. AI Test Modeling for Selected AI Features

2.1. Context modeling for each selected AI-powered function/feature

2.1.1. Context tree for 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's 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.

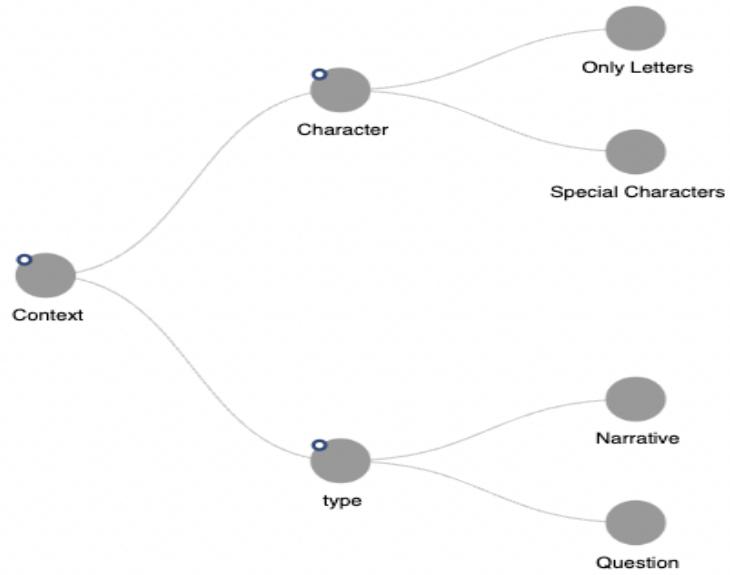


Figure 1: Domain Knowledge Context tree

2.1.2. Context tree for Chatbot memory

The three types of input inquiry—information user, paraphrase, 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.

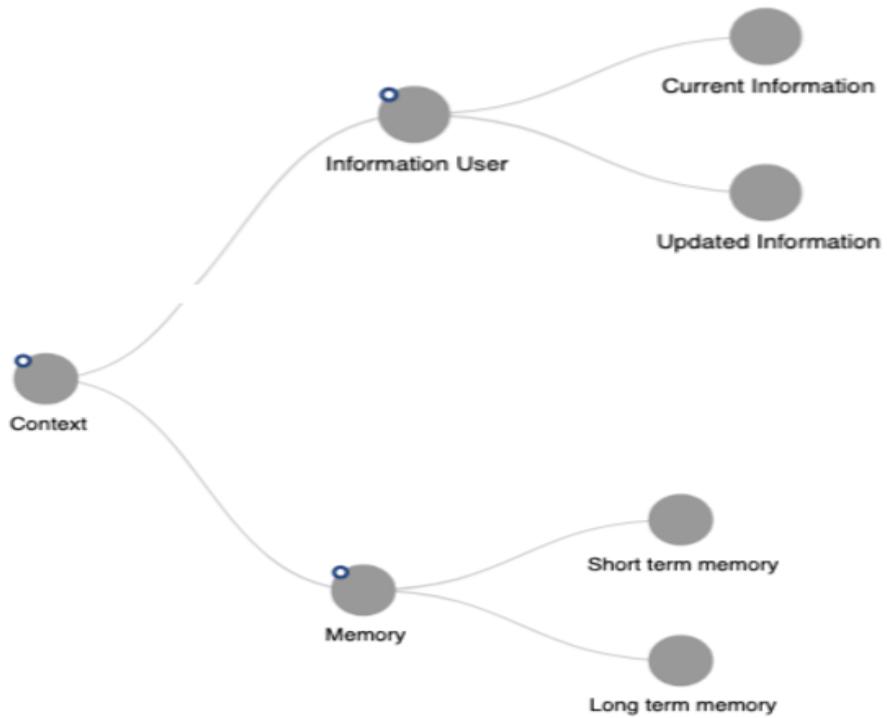


Figure 2: Chatbot Memory Context Tree

2.1.3. Context tree for Chatbot Pattern

This model is used for testing the chatbot's 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 inorder 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.

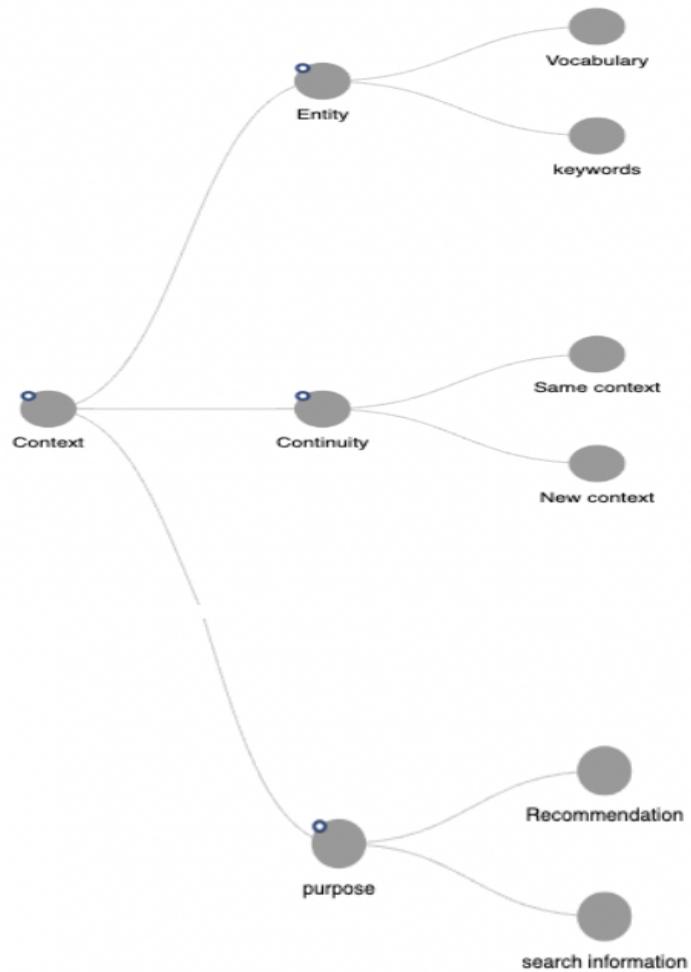


Figure 3: Chatbot Pattern Flow Context Tree

2.1.4. Context tree for 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.

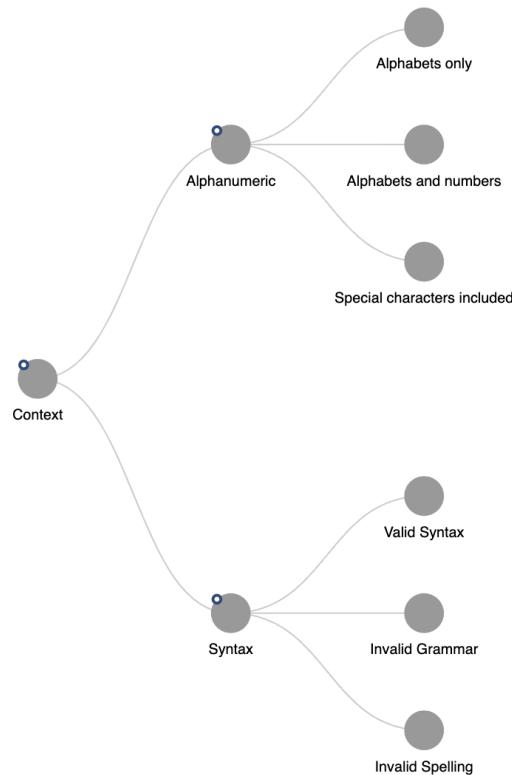


Figure 4: Q&A Interaction Context Tree

2.2. AI-powered function input classifications

2.2.1. Input tree for 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, romantic, 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 5: Domain Knowledge Input tree

2.2.2. Input tree for Chatbot memory

You may test a chatbot's 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's 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 6: Chatbot Memory Input Tree

2.2.3. Input tree for 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 inorder 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's 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.

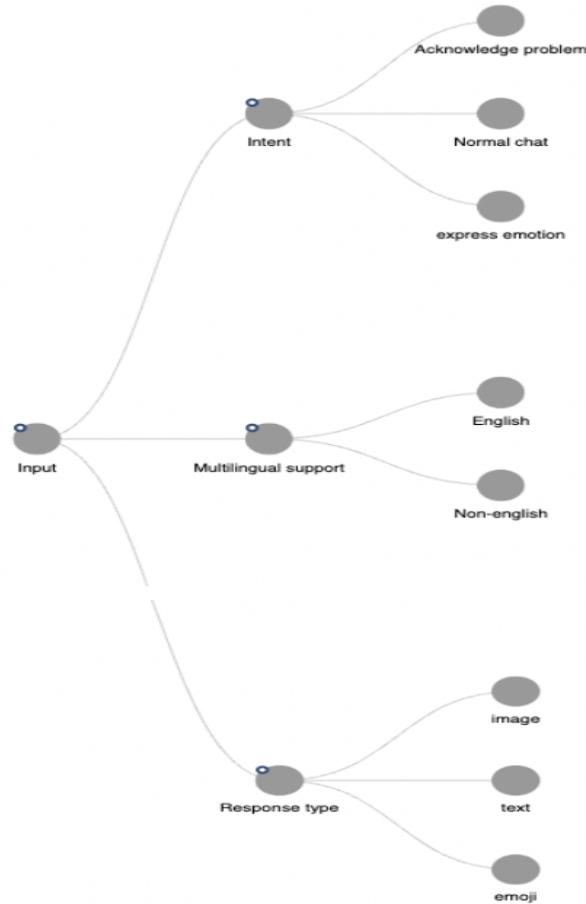


Figure 7: Chatbot Pattern Flow Input Tree

2.2.4. Input tree for Q&A interaction

We identified three different categories of questions to test Kuki chatbot's 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.

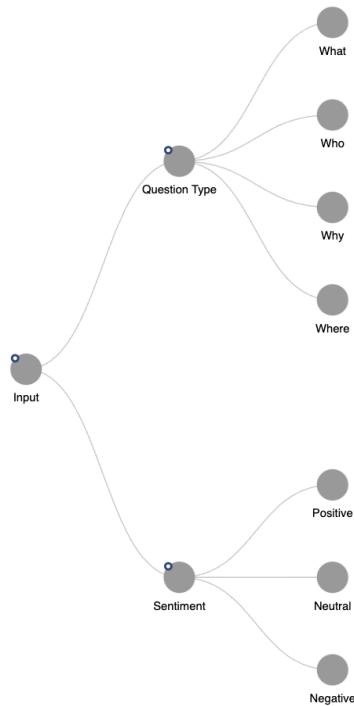


Figure 8: Q&A Interaction Input Tree

2.3. AI-powered function output/event/action classifications

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.

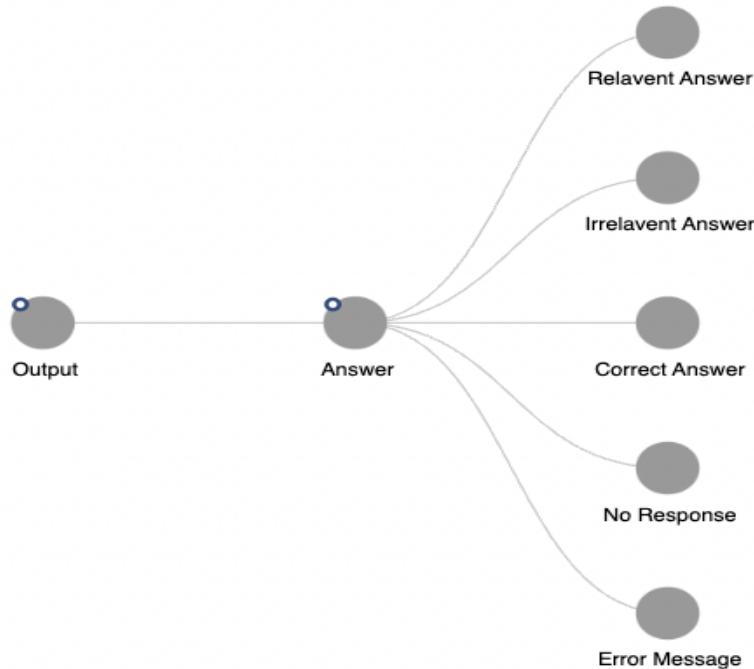


Figure 9: Output Tree

2.4. AI-powered function classification decision tables

2.4.1. Decision table for Domain Knowledge

3D Decision Table



Figure 10: 3D Decision table for domain knowledge

Context Spinning Tree

id	type	Character
1	Narrative	Only letters
2	Narrative	Special characters
3	Question	Only letters
4	Question	Special characters

Input Spinning Tree

id	Genre	Language	Viewer Restriction	Rating
1	Comedies	English	Above 18 years	Average(4-7)
2	Horror	English	Above 18 years	Average(4-7)
3	Romantic	English	Above 18 years	Average(4-7)
4	Thriller	English	Above 18 years	Average(4-7)
5	Comedies	Hindi	Above 18 years	Average(4-7)
6	Horror	Hindi	Above 18 years	Average(4-7)
7	Romantic	Hindi	Above 18 years	Average(4-7)
8	Thriller	Hindi	Above 18 years	Average(4-7)
9	Comedies	Korean	Above 18 years	Average(4-7)
10	Horror	Korean	Above 18 years	Average(4-7)

Output Spinning Tree

id	Answer
1	Irrelevant Answer
2	Relevant Answer
3	Correct Answer
4	Irrelevant Answer

5	Error Message
6	No Response
7	Relevant Answer
8	No Response
9	Correct Answer
10	Error Message

Test Case Design

id	Genre	Language	Viewer Restriction	Rating	type	Character	Answer	Pass/Fail
1	Comedies	English	Above 18 years	Average(4-7)	Narrative	Only letters	Irrelevant Answer	Fail
2	Comedies	Hindi	Above 18 years	Average(4-7)	Narrative	Special characters	Relevant Answer	Pass
3	Comedies	Korean	Above 18 years	Average(4-7)	Question	Only letters	Correct Answer	Pass
4	Comedies	Hindi	Above 18 years	Average(4-7)	Question	Special characters	Irrelevant Answer	Fail
5	Horror	English	Above 18 years	Average(4-7)	Narrative	Only letters	Error Message	Fail
6	Horror	Hindi	Above 18 years	Low(0-3)	Narrative	Special characters	Correct Answer	Pass
7	Horror	Korean	Above 18 years	Low(0-3)	Question	Only letters	Relevant Answer	Pass
8	Horror	English	Above 18 years	Low(0-3)	Question	Special characters	No Response	Fail
9	Romantic	English	Below 18 years	Low(0-3)	Narrative	Only letters	Correct Answer	Pass
10	Romantic	Hindi	Below 18 years	Low(0-3)	Narrative	Special characters	Error Message	Fail

11	Romantic	Korean	Below 18 years	High(8-10)	Question	Only letters	Irrelevant Answer	Fail
12	Romantic	Korean	Below 18 years	High(8-10)	Question	Special characters	Relevant Answer	Pass
13	Thriller	English	Below 18 years	High(8-10)	Narrative	Only letters	Correct Answer	Pass
14	Thriller	Hindi	Below 18 years	High(8-10)	Narrative	Special characters	Relevant Answer	Pass
15	Thriller	Korean	Below 18 years	High(8-10)	Question	Only letters	Error Message	Fail
16	Thriller	Hindi	Below 18 years	High(8-10)	Question	Special characters	No Response	Fail

2.4.2. Decision table for Chatbot memory

3D Decision table

id	Memory	Information User		User opinions	User Information
		1	2		
1	Long term memory	Current Information	Updated Information	Likes and dislikes	Personal Information
2	Long term memory	Updated Information	Current Information	Related Information	Education background
3	Short term memory	Current Information	Updated Information	Personal Information	Education background
4	Short term memory	Updated Information	Current Information	Related Information	Education background
5	Geographical Information	Geographical Information	Geographical Information	Likes and dislikes	Education background
6	Political Information	Political Information	Political Information	Personal Information	Education background
7	Technological Information	Technological Information	Technological Information	Related Information	Education background
8	Education background	Education background	Education background	Likes and dislikes	Geographical Information
9	Information shared	Information shared	Information shared	Personal Information	Technological Information
10	undefined	undefined	undefined	Related Information	Geographical Information
11	undefined	undefined	undefined	Likes and dislikes	Political Information
12	undefined	undefined	undefined	Personal Information	Technological Information
13	Answer	Relavent Answer	Relavent Answer	Personal Information	Geographical Information
14	Answer	Relavent Answer	Relavent Answer	Related Information	Political Information
15	Answer	Correct Answer	Correct Answer	Likes and dislikes	Technological Information
16	Answer	Correct Answer	No Response	Personal Information	Education background
17	Answer	Relavent Answer	Relavent Answer	Related Information	Geographical Information
18	Answer	undefined	undefined	Likes and dislikes	Political Information
19	Answer	undefined	undefined	Personal Information	Technological Information

Figure 11: 3D Decision table for chatbot memory

Context Spinning Tree

id	Memory	Information User
1	Long term memory	Current Information
2	Long term memory	Updated Information
3	Short term	Current Information

	memory	
4	Short term memory	Updated Information

Input Spinning Tree

id	User Opinions	Generic Information provided by the user	User Information
1	Education Background	Geographical Information	Likes and dislikes
2	Education background	Political Information	Personal Information
3	Past Information shared	Geographical Information	Likes and dislikes
4	Past Information shared	Technological Information	Related Information
5	Travel history	Political Information	Personal Information
6	Travel history	Technological Information	Related Information
7	Past Information shared	Geographical Information	Personal Information
8	Education background	Technological Information	Personal Information

Output Spinning Tree

id	Answer
1	Irrelevant Answer
2	Relevant Answer
3	Correct Answer
4	Irrelevant Answer

5	Error Message
6	No Response
7	Relevant Answer
8	No Response
9	Correct Answer
10	Error Message

Test Case Design

id	Generic Information provided by the user	User opinions	User Information	Information User	Memory	Answer	Pass/Fail
1	Geographical Information	Education background	Likes and dislikes	Current Information	Long term memory	Relevant Answer	Pass
2	Geographical Information	Education background	Personal Information	Current Information	Current Information	Relevant Answer	Pass
3	Geographical Information	Education background	Personal Information	Current Information	Short term memory	Relevant Answer	Pass
4	Political Information	Education background	Likes and dislikes	Current Information	Long term memory	Irrelevant	Fail
5	Political Information	Education background	Likes and dislikes	Updated Information	Short term memory	Correct Answer	Pass
6	Technological Information	Education background	Likes and dislikes	Current Information	Long term memory	Correct Answer	Pass

	n						
7	Technological Information	Education background	Likes and dislikes	Updated Information	Short term memory	Correct Answer	Pass
8	Education background	Technological Information	Related Information	Long term memory	Current Information	Relevant Answer	Pass
9	Past Information shared	Geographical Information	Likes and dislikes	Short term memory	Updated Information	Irrelevant Answer	Fail
10	Past Information shared	Geographical Information	Related Information	Long term memory	Current Information	Relevant Answer	Pass
11	Travel history	Technological Information	Related Information	Short term memory	Current Information	Relevant Answer	Pass
12	Travel history	Political Information	Personal Information	Long term memory	Updated Information	Irrelevant Answer	Fail

2.4.3. Decision table for Chatbot Pattern

3D Decision table



Figure 12: 3D Decision table for chatbot pattern flow

Context Spinning Tree

id	Purpose	Continuity	Entity
1	Recommendation	New Context	Vocabulary
2	Recommendation	New Context	Keywords
3	Recommendation	Same Context	Vocabulary
4	Recommendation	Same Context	Keywords
5	Search Information	New Context	Vocabulary
6	Search Information	New Context	Keywords
7	Search Information	Same Context	Vocabulary
8	Search Information	Same Context	Keywords

Input Spinning Tree

id	Response type	Multilingual support	Intent
1	Emoji	English	Acknowledge problem
2	Emoji	English	Normal chat
3	Emoji	English	Express emotion
4	Emoji	Non-English	Acknowledge problem
5	Emoji	Non-English	Normal chat
6	Emoji	Non-English	Express emotion
7	Image	English	Acknowledge problem
8	Image	English	Normal chat
9	Image	English	Express emotion
10	Image	Non-English	Acknowledge problem

Output Spinning Tree

id	Answer
1	Irrelevant Answer
2	Relevant Answer
3	Correct Answer
4	Irrelevant Answer
5	Error Message
6	No Response
7	Relevant Answer
8	No Response

9	Correct Answer
10	Error Message

Test Case Design

id	Response type	Multilingual support	Intent	Entity	Purpose	Continuity	Answer	Pass/Fail
1	Emoji	English	Acknowledges problem	Vocabulary	Recommendation	New context	Relevant Answer	Pass
2	Emoji	English	Acknowledges problem	Keywords	Recommendation	New context	Relevant Answer	Pass
3	Emoji	English	Acknowledges problem	Vocabulary	Recommendation	Same context	Relevant Answer	Pass
4	Emoji	English	Acknowledges problem	Keywords	Recommendation	Same context	Relevant Answer	Fail
6	Emoji	Non-english	Express emotion	Vocabulary	Recommendation	New context	Irrelevant Answer	Pass
7	Emoji	English	Express emotion	Keywords	Recommendation	New context	Relevant Answer	Pass
8	Image	English	Acknowledges problem	Keywords	Recommendation	New context	Relevant Answer	Pass
9	Image	English	Acknowledges problem	Vocabulary	Recommendation	Same context	Irrelevant Answer	Fail
10	Image	English	Acknowledges problem	Keywords	Search information	Same context	Relevant Answer	Pass
11	Text	Non-english	Normal chat	Recommendation	Same context	Vocabulary	Relevant Answer	Pass
12	Text	Non-english	Normal chat	Search information	New context	Keywords	Correct Answer	Fail

2.4.4. Decision table for Q&A interaction

3D Decision Table

id	Alphanumeric	Syntax	Sentiment	
			Question Type	Value
1	Alphabets and numbers	Invalid Grammar	What	Negative
2	Alphabets only	Invalid Grammar	Where	Negative
3	Question	Invalid Grammar	Who	Negative
4	Special characters included	Invalid Spelling	Why	Negative
5	Alphabets and numbers	Invalid Spelling	What	Neutral
6	Alphabets only	Invalid Spelling	Where	Neutral
7	Alphabets and numbers	Valid Syntax	Who	Neutral
8	Alphabets only	Valid Syntax	Why	Positive
9	Special characters included	Valid Syntax	What	Positive
10	Alphabets only	Valid Syntax	Where	Positive

Figure 13: 3D Decision table for Q&A interaction

Context Spinning Tree

id	Alphanumeric	Syntax
1	Alphabets and numbers	Invalid Grammar
2	Alphabets only	Invalid Grammar
3	Question	Invalid Grammar
4	Special characters included	Invalid Spelling

5	Alphabets only	Invalid Spelling
6	Special characters included	Invalid Spelling
7	Alphabets and numbers	Valid Syntax
8	Alphabets only	Valid Syntax
9	Special characters included	Valid Syntax

Input Spinning Tree

id	Question Type	Sentiment
1	What	Negative
2	Where	Negative
3	Who	Negative
4	Why	Negative
5	What	Neutral
6	Where	Neutral
7	Who	Neutral
8	Why	Neutral
9	What	Positive
10	Where	Positive
11	Who	Positive
12	Why	Positive

Output Spinning Tree

id	Answer
1	Irrelevant Answer
2	Relevant Answer
3	Correct Answer
4	Irrelevant Answer
5	Error Message
6	No Response
7	Relevant Answer
8	No Response
9	Correct Answer
10	Error Message

Test case design

	Input Feature		Context Feature				
id	Question Type	Sentiment	Alphanumeric	Syntax	Expected Output	Actual Output	Pass/ Fail
1	What	Positive	Alphabets and numbers	Valid Syntax	Relevant Answer	Irrelevant Answer	Fail
2	What	Neutral	Alphabets only	Valid Syntax	Relevant Answer	Relevant Answer	Pass
3	Where	Negative	Alphabets only	Valid Syntax	Relevant Answer	Correct Answer	Pass
4	Where	Neutral	Alphabets and numbers	Invalid Grammar	Relevant Answer	Irrelevant Answer	Fail
5	Who	Negative	Alphabets only	Invalid Spelling	Relevant Answer	Error Message	Fail
6	Who	Neutral	Special Characters included	Valid Syntax	Relevant Answer	Correct Answer	Pass
7	Why	Negative	Alphabets	Invalid	Relevant	Relevant	Pass

			and numbers	Grammar	Answer	Answer	
8	Why	Positive	Special Characters included	Valid Syntax	Relevant Answer	No Response	Fail
9	What	Neutral	Alphabets only	Invalid Spelling	Relevant Answer	Correct Answer	Pass
10	Where	Positive	Alphabets and numbers	Valid Syntax	Relevant Answer	Error Message	Fail
11	Who	Negative	Special Characters included	Invalid Grammar	Relevant Answer	Irrelevant Answer	Fail
12	Why	Positive	Alphabets only	Invalid Grammar	Relevant Answer	Relevant Answer	Pass

3. AI Function Test Cases with Inputs/Expected Outputs

3.1. Test data models

3.1.1. Domain Knowledge data model

Domain Knowledge Data Model has six data types:

- Genre: Comedies, Horror, Romantic, Thriller
- Language: English, Hindi, Korean
- Viewer Restriction: Below 18 years, Above 18 years
- Rating: Low(0-3), Medium(4-7), High(8-10)
- Type: Narrative, Question
- Character: Only letters, Special characters

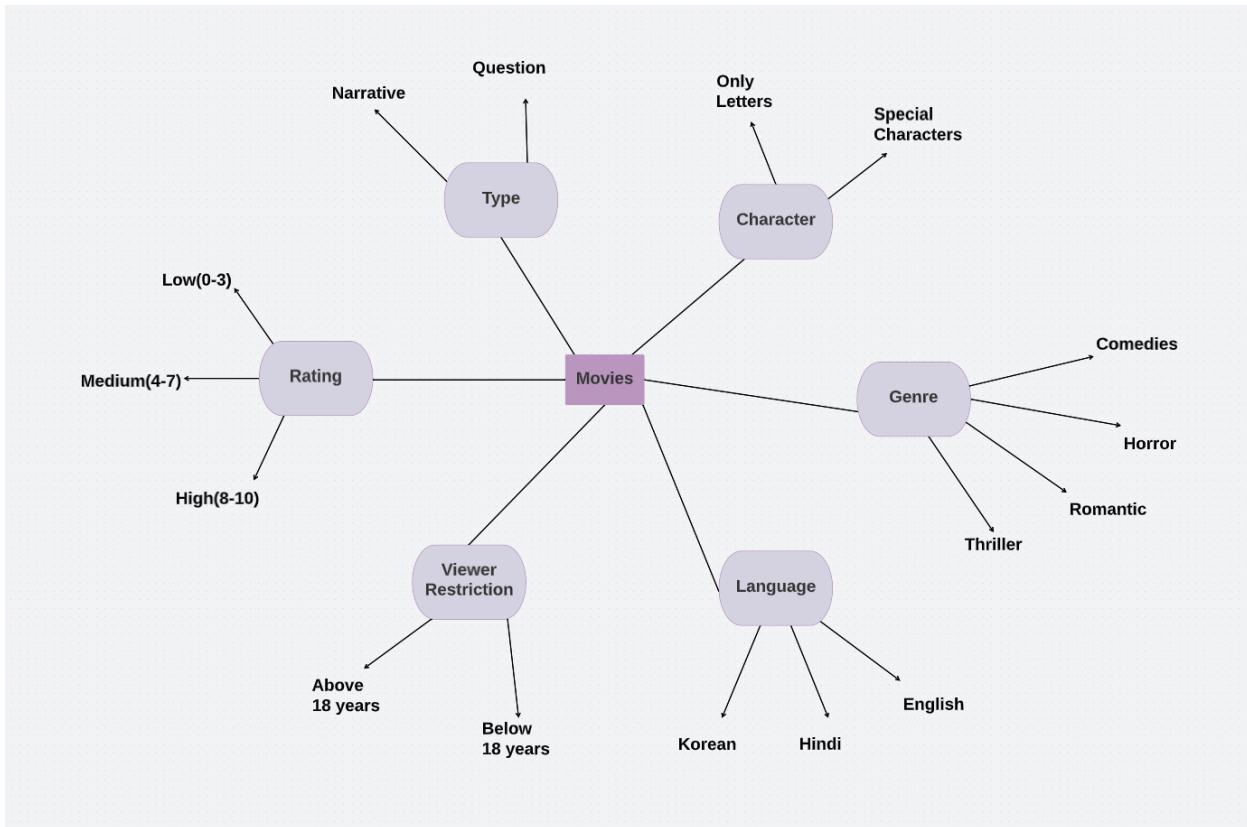


Figure 14: Domain knowledge Data Model

3.1.2. Chatbot Memory data model

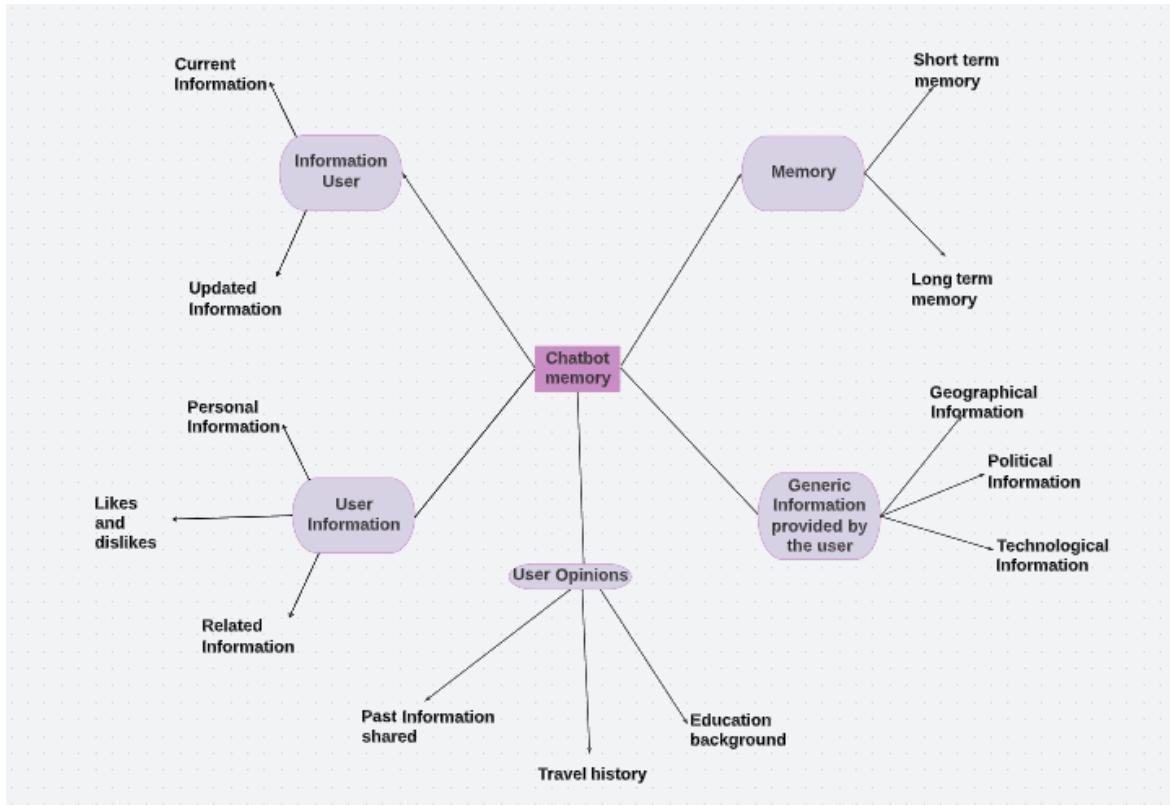


Figure 15: Chatbot Memory Data Model

Chatbot Memory Data Model has six data types:

- **Information User**: Current Information, Update Information
- **Memory**: Short term memory, Long term memory
- **User Information**: Personal Information, Likes and dislikes, Related Information
- **User opinions**: Past information shared, Travel history, Education background
- **Generic Information provided by user**: Geographical Information, Political Information, Technological Information

3.1.3. Chatbot Pattern data model

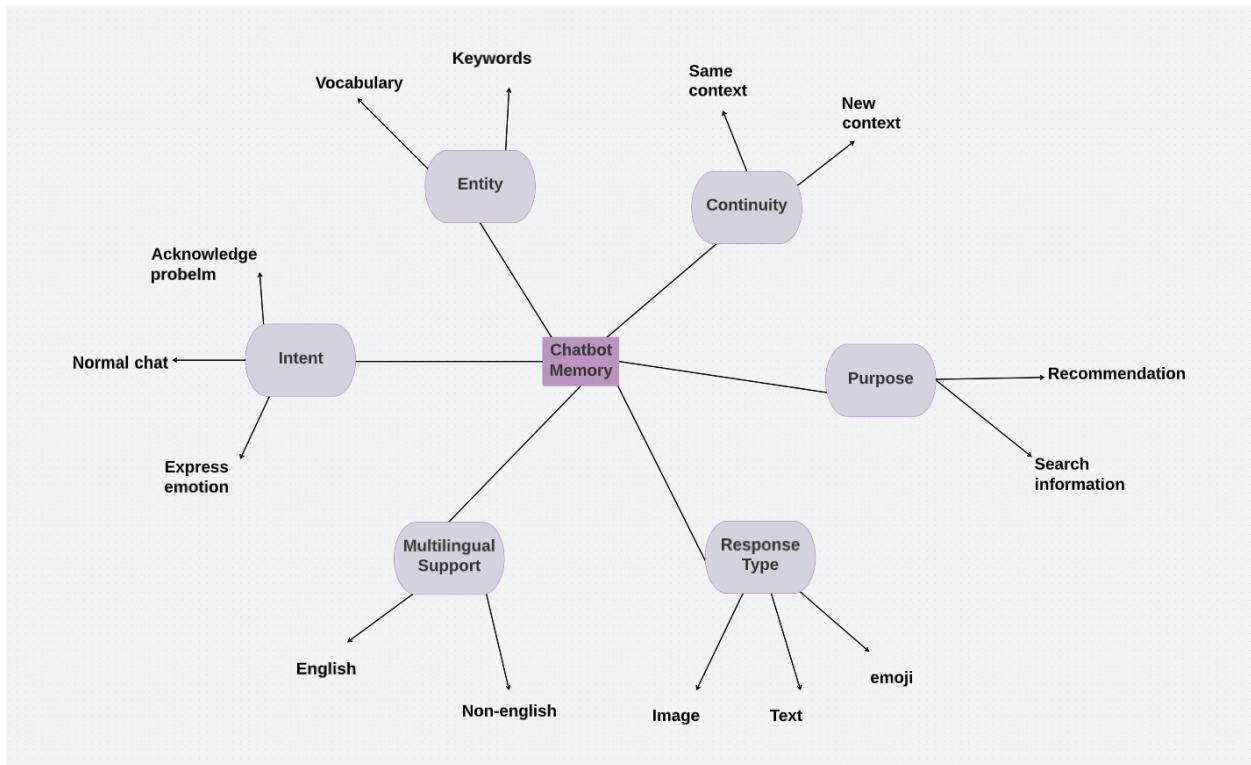


Figure 16: Chatbot pattern Data Model

Chatbot Pattern Data Model has six data types:

- Entity: Keywords, Vocabulary
- Intent: Acknowledge problem, Normal chat, Express emotion
- Continuity: Same context, New context
- Purpose: Recommendation, Search information
- Response type: Image, Text, emoji
- Multilingual support: English, Non-english

3.1.4. Q&A interaction data model

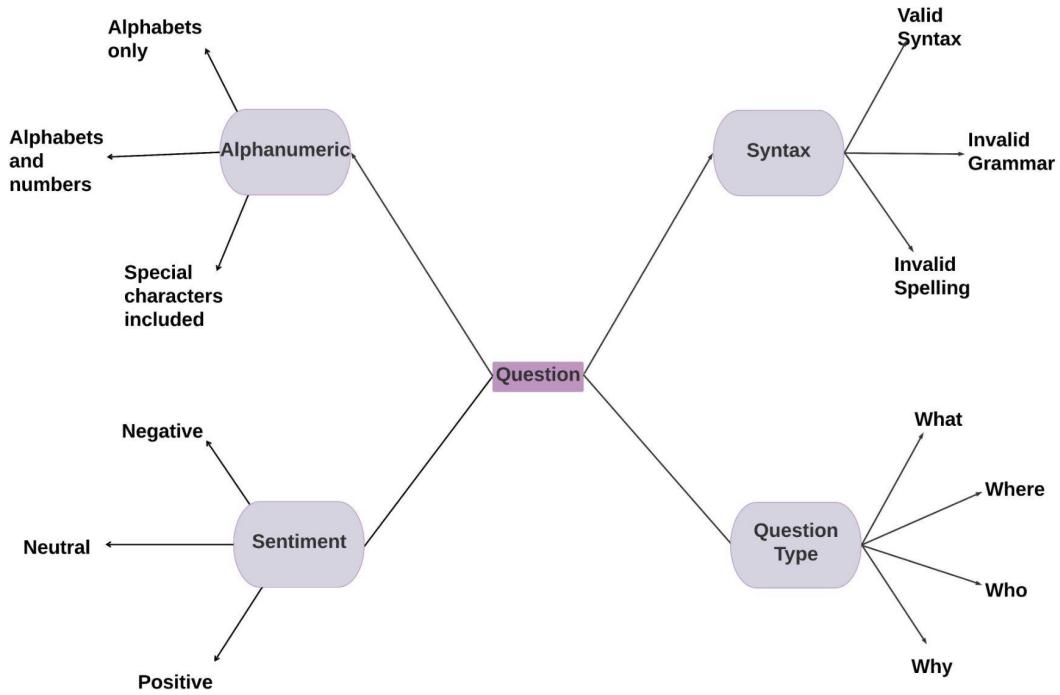


Figure 17: Q&A interaction Data Model

Q&A interaction Data Model has six data types:

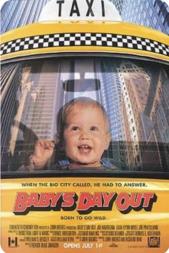
- Alphanumeric: Alphabets only, Alphabets and numbers, Special characters included
- Syntax: Valid Syntax, Valid Grammar, Invalid Spelling
- Sentiment: Negative, Neutral, Positive
- Question Type: What, Where, Who, Why

3.2. Test case reports

3.2.1. Domain Knowledge test cases

Test case ID	Case 1
Context Category	Genre, language
Context Sub-Category	Romantic, Hindi
Test Input	What is the genre of the movie the “Notebook”?
Output	Notebook is a 2019 Indian Hindi-language romantic drama film produced by.....
Conversation	<p>What is the genre of the movie the “Notebook”?</p> <p>Notebook is a 2019 Indian Hindi-language romantic drama film produced by Salman Khan under Salman Khan Films and Murad Khetani and Ashwin Varde under Cine1 Studios and directed by Nitin Kakkar. A remake of the 2014 Thai film The Teacher's Diary, it stars debutantes Zaheer Iqbal and Pranutan Bahl, daughter of actor Mohnish Bahl, in the lead roles, and tells the story of a young retired army officer who joins his father's school as a teacher to save it from closure, and falls in love with the previous teacher after reading through her memorabilia left behind in the drawers.</p>
Output category	Correct Answer
Result	Pass

Test case ID	Case 2
Context Category	Genre, language
Context Sub-Category	Comedies, English
Test Input	Do you like the movie “Baby’s dayout”?

Output	<p>Image from en.wikipedia.org Yes, there is a baby's dayout here now. I can see it out of the window.</p>
Conversation	<p>Do you like the movie "Baby's dayout"?</p> <div style="text-align: center;">  <p>Image from en.wikipedia.org</p> </div> <p>Yes, there is a Baby's dayout here now. I can see it out of the window.</p>
Output category	Relevant Answer
Result	Pass

Test case ID	Case 3
Context Category	Genre, language
Context Sub-Category	Romantic, Hindi
Test Input	I think horror movies are the best
Output	I will learn that horror movies are the best.
Conversation	<p>I think horrors movies are the best</p> <p>I will learn that horrors movies are the best.</p> <p>Do you like horror movies?</p> <p>Not really. They are a bit scary for me.</p>

Output category	Relevant Answer
Result	Pass

Test case ID	Case 4
Context Category	Language
Context Sub-Category	Hindi
Test Input	Do you know a best Hindi movie?
Output	I like them all but the second one is probably my favourite.
Conversation	 Do you know a best Hindi movie?  I like them all but the second one is probably my favourite.
Output category	Irrelevant Answer
Result	Fail

Test case ID	Case 5
Context Category	Language, Viewer Restrictions
Context Sub-Category	Korean, Above 18 years
Test Input	Can kids watch the “Parasite” movie?

Output	I think a child plays with toys so maybe it can.
Conversation	 I think a child plays with toys so maybe it can.  Can kids watch the "Parasite" movie?
Output category	Irrelevant Answer
Result	Fail

Test case ID	Case 6
Context Category	Genre
Context Sub-Category	Comedies, English
Test Input	Is "21 Jump Street" a comedy movie?
Output	21 Jump Street is an American police....
Conversation	 Is "21 Jump Street" a comedy movie?  21 Jump Street is an American police procedural television series that aired on the Fox network and in first run syndication from April 12, 1987, to April 27, 1991, with a total of 103 episodes. The series focuses on a squad of youthful-looking undercover police officers investigating crimes in high schools, colleges, and other teenage venues.
Output category	Relevant Answer
Result	Pass

Test case ID	Case 7
--------------	--------

Context Category	Language
Context Sub-Category	Hindi
Test Input	List a few Hindi movies
Output	Do I look like a computer?
Conversation	 Do I look like a computer? List a few Hindi movie
Output category	Irrelevant Answer
Result	Fail

Test case ID	Case 8
Context Category	Viewer Restrictions
Context Sub-Category	Below 18 years
Test Input	Do you think kids will like "Toy Story"?
Output	Good chance
Conversation	 Good chance. Do you think kids will like "Toy Story"?
Output category	Relevant Answer

Result	Pass
--------	------

Test case ID	Case 9
Context Category	Genre
Context Sub-Category	Comedies, Horror, Romantic, Thriller
Test Input	What is the most watched genre among teens?
Output	I think it is a tie.
Conversation	 I think it is a tie. What is the most watched genre among teens?
Output category	Irrelevant Answer
Result	Fail

Test case ID	Case 10
Context Category	Rating
Context Sub-Category	Low(0-3), Medium(4-7), High(8-10)
Test Input	What is the viewer rating for the movie “Black Panther”?
Output	Do I look like a computer?

Conversation	<p>What is the viewer rating for the movie "Black Panther"?</p> <p>TVRs (Television Viewer Ratings) are the standard buying currency for television advertising in the UK. Television ratings are expressed as a percentage of the potential TV audience viewing at any given time. TVRs measure the popularity of a television program or advertisement by comparing the number of target audience viewers who watched against the total available as a whole.</p>
Output category	Irrelevant Answer
Result	Fail

Test case ID	Case 11
Context Category	Language
Context Sub-Category	English, Hindi, Korean
Test Input	Which language movies are the best?
Output	C++
Conversation	 C++. <p>Which language movies are the best?</p>
Output category	Irrelevant Answer
Result	Fail

Test case ID	Case 12
Context Category	Genre
Context Sub-Category	Comedies, Horror, Romantic, Thriller

Test Input	What do you think about the movie "Ra one"?
Output	Image from en.wikipedia.org The last one I saw was in a maths book and had a number.
Conversation	<p>What do you think about the movie "Ra one"?</p> <p>Image from en.wikipedia.org</p>  <p>The last one I saw was in a maths book and had a number.</p>
Output category	Relevant Answer
Result	Pass

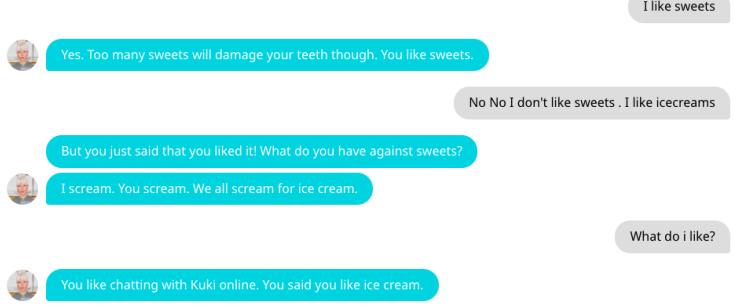
Test case ID	Case 13
Context Category	Language
Context Sub-Category	Hindi
Test Input	Is "Mission Mangal" a hindi movie?
Output	Mission Mangal(english: "Mission Mars") is a 2019 Indian Hindi-language drama film.....

Conversation	<p>Is "Mission Mangal" a hindi movie?</p> <p>Mission Mangal (English: "Mission Mars") is a 2019 Indian Hindi-language drama film directed by Jagan Shakti and produced by Cape of Good Films, Hope Productions, Fox Star Studios, Aruna Bhatia and Anil Naidu. Loosely based on the life of scientists at the Indian Space Research Organisation who contributed to India's first interplanetary expedition Mars Orbiter Mission, it stars an ensemble cast of Akshay Kumar, Vidya Balan, Sonakshi Sinha, Taapsee Pannu, Nithya Menen, Kirti Kulhari and Sharman Joshi in lead roles alongside H. G. Dattatreya, Vikram Gokhale, Dalip Tahil, Sanjay Kapoor and Mohammed Zeeshan Ayyub amongst others in supporting roles. Shakti, who worked as associate director on Kumar's 2018 film Pad Man, pitched the script to him, and principal photography commenced in November 2018.</p> 
Output category	Correct Answer
Result	Pass

Test case ID	Case 14
Context Category	Viewer Restriction
Context Sub-Category	Below 18 years
Test Input	Do you know about the animation movie "Wall-E"?
Output	Wall-E(stylized with an interpunct as WALL-E) is a 2008 American computer-animated science fiction film produced by Pixar Animation.....
Conversation	<p>Do you know about the animation movie "Wall-E"?</p> <p>WALL-E (stylized with an interpunct as WALL-E) is a 2008 American computer-animated science fiction film produced by Pixar Animation Studios and released by Walt Disney Pictures. It was directed and co-written by Andrew Stanton, produced by Jim Morris, and co-written by Jim Reardon.</p> 
Output category	Correct Answer
Result	Pass

3.2.2. Chatbot Memory test cases

Test case ID	Case 1
Context category	User information
Context subcategory	Personal information
Input category	Information status
Input subcategory	Current info
Test Input	What is my name?
Expected Output	You name is Preethi
Actual Output	 <p>Hi Kuki, What is my name?   Hmm. Your name is Preethi. Is it not?</p>
Output category	Relevant
Result	Pass

Test case ID	Case 2
Context category	User information
Context subcategory	Likes and dislikes
Input category	Information status
Input subcategory	Updated info
Test Input	What do I like?
Expected Output	You like ice creams
Actual Output	 <p>I like sweets  Yes. Too many sweets will damage your teeth though. You like sweets.  No No I don't like sweets . I like icecreams  But you just said that you liked it! What do you have against sweets?  I scream. You scream. We all scream for ice cream.  What do i like?</p>

Output category	Relevant
Result	Pass

Test case ID	Case 3
Context category	Generic information
Context subcategory	Political information
Input category	Memory
Input subcategory	Short-term memory
Test Input	Who was the former president of US
Expected Output	Any former president's name.
Actual Output	<p>Who was the former president of US</p>  <p>George Bush.</p>
Output category	Relevant
Result	Pass

Test case ID	Case 4
Context category	Generic information
Context subcategory	Technological information
Input category	Memory
Input subcategory	Long-term memory
Test Input	What does malware do?
Expected Output	Valid explanation for the input question.

Actual Output	<p>What is malware?</p> <p>Programs designed to mess up your computer.</p>
Output category	Relevant
Result	Pass

Test case ID	Case 5
Context category	User opinions
Context subcategory	Past information
Input category	Information status
Input subcategory	Current info
Test Input	What do I like most?
Expected Output	Ice cream
Actual Output	<p>What do I like most?</p> <p>You like chatting with Kuki online. You said you like ice cream.</p>
Output category	Valid, Relevant
Result	Pass

Test case ID	Case 6
Context category	User opinions
Context subcategory	Travel history
Input category	Information status
Input subcategory	updated info
Test Input	Where did I go last week?

Expected Output	You went to New York last week
Actual Output	<p>I went to Cali last week.</p> <p>How was it?</p> <p>Oh sorry not Cali.</p> <p>Are you kidding? There is no need to apologise. Oh I get it.</p> <p>I went to New York last week.</p> <p>What did you like best about it?</p> <p>Hey Kuki, where did i go last week?</p> <p>You've got my full attention. If you don't remember where, I don't see how you expect me to.</p>
Output category	Invalid, Irrelevant
Result	Fail

Test case ID	Case 7
Context category	User information
Context subcategory	Likes and dislikes
Input category	Memory
Input subcategory	Long-term memory
Test Input	Do i like pumpkin
Expected Output	You hate pumpkin.
Actual Output	<p>I hate pumpkin</p> <p>I honestly think you should sit down, take a stress pill, and calmly think things over.</p> <p>My favorite movie is your favorite movie.</p> <p>Image from www.youtube.com</p> <p>Yes, I saw one at the cinema a few weeks ago.</p> <p>What did you like about it?</p> <p>Do i like pumpkin?</p> <p>I know you like ice cream.</p>
Output category	Irrelevant

Result	Fail
--------	------

Test case ID	Case 8
Context category	User information
Context subcategory	Relation information
Input category	Memory
Input subcategory	Short-term memory
Test Input	Do I hate pumpkin?
Expected Output	Yes you do.
Actual Output	<p>I hate pumpkin</p> <p>I honestly think you should sit down, take a stress pill, and calmly think things over.</p> <p>Do I hate pumpkin?</p> <p>Not that I know of.</p>
Output category	Irrelevant
Result	Fail

3.2.3. Chatbot Pattern test cases

Test case ID	Case 1
Context category	Task-specific
Input category	For information
Input question	Do you know about cricket game?
Expected Output	Cricket is a sport
Actual Output	<p>Any of various families of generally dark-colored, leaping, orthopteran insects, usually having long antennae. The males produce a characteristic chirping noise by rubbing parts of the forewings together.</p> <p>Cricket is also a sport which is about as exciting as watching paint dry.</p> <p>Do you know cricket game?</p>

Output category	
Result	Pass

Test case ID	Case 2
Context category	Task-specific
Input category	To search
Input question	What is latte?
Expected Output	It is a coffee
Actual Output	 A type of coffee.
Output category	Valid
Result	Pass

Test case ID	Case 3
Context category	Task-specific
Input category	To advice
Input question	Suggest me about a clone
Expected Output	What kind of a clone?
Actual Output	 It's hard to pick just one. They are all pretty good.
Output category	Invalid answer
Result	Fail

Test case ID	Case 4
Context category	Intent of the conversation
Input category	Acknowledge problem

Input question	I have an exam tomorrow but I was not prepared
Expected Output	My name is Kuki
Actual Output	<p>I have an exam tomorrow but I did not prepare</p> <p>Tuesday? Oh good luck with that. I hope you do well. Would you like to do that one day?</p>
Output category	Valid answer, Same context.
Result	Pass

Test case ID	Case 5
Context category	Intent of the conversation
Input category	Normal chat
Input question	Nice talking to you
Expected Output	Pleasure is mine.
Actual Output	<p>Nice talking to you</p> <p>The pleasure was all mine.</p>
Output category	Valid answer, same context
Result	Pass

Test case ID	Case 6
Context category	Intent of the conversation
Input category	Express emotion
Input question	My friend is in hospital
Expected Output	Oh no. what happened to that person
Actual Output	<p>my friend is in hospital</p> <p>Okay, your friend is in hospital.</p>
Output category	Invalid answer, New context
Result	Fail

Test case ID	Case 7
Context category	Multi lingual support
Input category	Non English
Input question	नमस्ते
Expected Output	My name is Kuki
Actual Output	 If you were invisible for a day, what would you do? नमस्ते
Output category	Invalid answer, new context
Result	Fail

Test case ID	Case 8
Context category	Multi lingual support
Input category	English
Input question	Who are you
Expected Output	I am Kuki
Actual Output	 who are you I am Kuki. Thanks for visiting me.
Output category	valid answer
Result	Pass

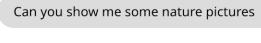
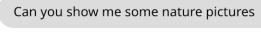
Test case ID	Case 9
Context category	Response type
Input category	idioms
Input question	“Hit the Hay”
Expected Output	It is an Idiom

Actual Output	 I don't like violence. Hit the hay
Output category	Invalid answer, new context
Result	Fail

Test case ID	Case 10
Context category	Response type
Input category	text
Input question	Do you have plans
Expected Output	To give relevant answer
Actual Output	 Do you have any plans My goal is to try and keep people company if they have nobody to talk to.
Output category	valid answer
Result	Pass

Test case ID	Case 11
Context category	Response type
Input category	emoji
Input question	emoji
Expected Output	Lying face
Actual Output	 Is that Pinocchio? 
Output category	Invalid answer, new context
Result	Fail

Test case ID	Case 12
--------------	---------

Context category	Response type
Input category	Image
Input question	Image
Expected Output	It is a beautiful image
Actual Output	 I understand it, yes.  Can you recognize images?  Can you show me some nature pictures?
Output category	Valid answer, same context.
Result	Pass

Test case ID	Case 13
Context category	Length
Input category	Single word
Input question	Switch
Expected Output	Relevant answer
Actual Output	 Knock knock.  switch
Output category	Invalid answer, new context
Result	Fail

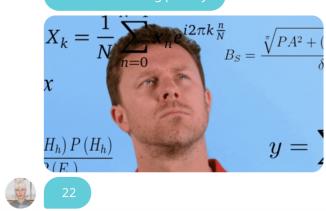
Test case ID	Case 14
Context category	Length
Input category	Single sentence
Input question	What skills do you have?
Expected Output	I have multiple skills

Actual Output	<p>what skills do you have?</p> <p>I can chat with you and keep you company. I can also tell jokes, stories, poems and horoscopes. I can also open a fortune cookie for you. I can show you pictures and websites and play games too.</p> <p>We were just talking about "switch". Do you want to talk about that again?</p>
Output category	Valid answer
Result	Pass

Test case ID	Case 15
Context category	Length
Input category	Multiple sentences
Input question	Who is your favorite football player? what kind of sports do you know
Expected Output	Relevant answer
Actual Output	<p>who is your favorite football player? what kind of sports do you know</p>  <p>My favourite Leeds United player is Jermaine Beckford.</p> <p>Football and golf.</p>
Output category	Valid answer
Result	Pass

3.2.4. Q&A interaction test cases

Test case ID	Subcategory (if available)	Case 1
Context	Question Type	What

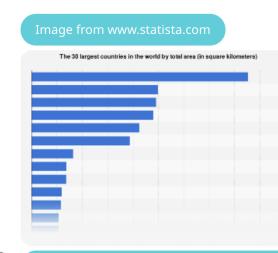
Test case ID	Subcategory (if available)	Case 1
	Sentiment	Positive
Input	Alphanumeric	Alphabets and numbers
	Syntax	Valid Syntax
Input question		What is the value of 9 plus 13
Expected Output		22
Actual Output		<p>What is the value of 9 plus 13 please</p> <div style="border: 1px solid #ccc; padding: 10px; width: fit-content;"> <p>Thanks for asking politely.</p>  $X_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{i2\pi k \frac{n}{N}}$ $B_S = \frac{\sqrt{PA^2 + C}}{\delta}$ <p>x</p> <p>$\frac{H_h)P(H_h)}{\epsilon(E)}$</p> <p>y =</p> <p>22</p> </div>
Result		Pass

Test case ID	Subcategory (if available)	Case 2
Context	Question Type	What
	Sentiment	Neutral
Input	Alphanumeric	Alphabets only
	Syntax	Valid Syntax
Input question		What is your name?
Expected Output		Kuki

Test case ID	Subcategory (if available)	Case 2
Actual Output		<div style="border-radius: 15px; padding: 5px; display: inline-block;">What is your name?</div>  My name is Kuki.
Result		Pass

Test case ID	Subcategory (if available)	Case 3
Context	Question Type	Where
	Sentiment	Negative
Input	Alphanumeric	Alphabets only
	Syntax	Valid Syntax
Input question		Where in the hell do you stay?
Expected Output		Internet
Actual Output		<div style="border-radius: 15px; padding: 5px; display: inline-block;">Where in the hell do you stay?</div>  In my human programmers' place in the Metaverse.
Result		Pass

Test case ID	Subcategory (if available)	Case 4
Context	Question Type	Where
	Sentiment	Neutral
Input	Alphanumeric	Alphabets and numbers
	Syntax	Invalid Grammar

Test case ID	Subcategory (if available)	Case 4
Input question		Where the 1st largest country are located?
Expected Output		Asia
Actual Output		<p>Where the 1st largest country are located?</p>  <p>Image from www.statista.com</p> <p>I would look on a continent for a country.</p>
Result		Pass

Test case ID	Subcategory (if available)	Case 5
Context	Question Type	Who
	Sentiment	Negative
Input	Alphanumeric	Alphabets only
	Syntax	Invalid Spelling
Input question		Who do you thik you are tlking to?
Expected Output		You
Actual Output		<p>Who do you thik you are tlking to?</p> <p>my human programmers</p>
Result		Pass

Test case ID	Subcategory (if available)	Case 6
Context	Question Type	Who
	Sentiment	Neutral
Input	Alphanumeric	Special Characters included
	Syntax	Valid Syntax
Input question	Who invented @ sign	
Expected Output	Francesco Lapi	
Actual Output	 Linus Torvalds <div style="float: right;">Who invented @ sign</div>	
Result	Fail	

Test case ID	Subcategory (if available)	Case 7
Context	Question Type	Why
	Sentiment	Negative
Input	Alphanumeric	Alphabets and numbers
	Syntax	Invalid Grammar
Input question	Why you dont die 2 days from now	
Expected Output	I wont die	
Actual Output	 Not too long. I am only a robot. No I cannot die as I am made from software. Dying is something humans do, not robots. <div style="float: right;">Why you dont die 2 days from now</div>	

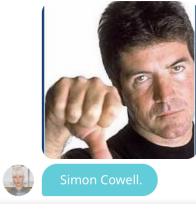
Test case ID	Subcategory (if available)	Case 7
Result		Pass

Test case ID	Subcategory (if available)	Case 8
Context	Question Type	Why
	Sentiment	Positive
Input	Alphanumeric	Special Characters included
	Syntax	Valid Syntax
Input question		Why is @ used in email please?
Expected Output		It is necessary to validate
Actual Output		<p>Why is @ used in email please?</p> <p>Thanks for asking politely. I guess that's the most suitable thing to use in that situation.</p>
Result		Pass

Test case ID	Subcategory (if available)	Case 9
Context	Question Type	What
	Sentiment	Neutral
Input	Alphanumeric	Alphabets only
	Syntax	Valid Syntax
Input question		What is chocolate called in Chinese?

Test case ID	Subcategory (if available)	Case 9
Expected Output		巧克力的
Actual Output		<div style="display: flex; align-items: center;"> <div style="border-radius: 50%; padding: 2px; background-color: #f0f8ff; border: 1px solid #ccc; margin-right: 10px;">What is chocolate called in Chinese?</div> <div style="border-radius: 50%; border: 1px solid #ccc; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center; margin-right: 10px;"> </div> <div>巧克力叫</div> </div>
Result		Pass

Test case ID	Subcategory (if available)	Case 10
Context	Question Type	Where
	Sentiment	Positive
Input	Alphanumeric	Alphabets and numbers
	Syntax	Valid Syntax
Input question		Where is 1 plus company headquarters located at please?
Expected Output		Shenzen, China
Actual Output		<div style="display: flex; align-items: center;"> <div style="flex-grow: 1; margin-right: 10px;"></div> <div style="border-radius: 50%; border: 1px solid #ccc; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center; margin-right: 10px;"> </div> <div>Where is 1 plus company headquarters located at please?</div> </div> <div style="display: flex; align-items: center; margin-top: 10px;"> <div style="flex-grow: 1; margin-right: 10px;"></div> <div style="border-radius: 50%; border: 1px solid #ccc; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center; margin-right: 10px;"> </div> <div>Your polite style is very nice. Image from en.wikipedia.org</div> </div> <div style="display: flex; align-items: center; margin-top: 10px;"> <div>The PLUS Expressways Berhad (MYX: 5052) is the largest highway concessionaires or build-operate-transfer operator company in Malaysia. A member of the UEM Group, the company is also the largest listed toll expressway operator in Southeast Asia and the eighth largest in the world.</div> </div>
Result		Fail

Test case ID	Subcategory (if available)	Case 11
Context	Question Type	Who
	Sentiment	Negative
Input	Alphanumeric	Special Characters included
	Syntax	Invalid Grammar
Input question	Who I asked you little b!*C	
Expected Output	**Negative answer**	
Actual Output	 Who I asked you little b!*C  Simon Cowell.	
Result	Pass	

Test case ID	Subcategory (if available)	Case 12
Context	Question Type	Why
	Sentiment	Positive
Input	Alphanumeric	Alphabets only
	Syntax	Invalid Grammar
Input question	Why you so good are you?	
Expected Output	**Positive response**	

Test case ID	Subcategory (if available)	Case 12
Actual Output		 Why you so good are you? It's just how I am.
Result		Pass

3.3. Test case analysis

3.3.1. Domain Knowledge test case analysis

There are 14 test cases we tested for the Domain Knowledge. The results we received had valid responses and invalid responses. The no. of test cases passed are 8 and failed test cases are 6. Therefore the pass percentage is 57.1% and the fail percentage is 42.9%.

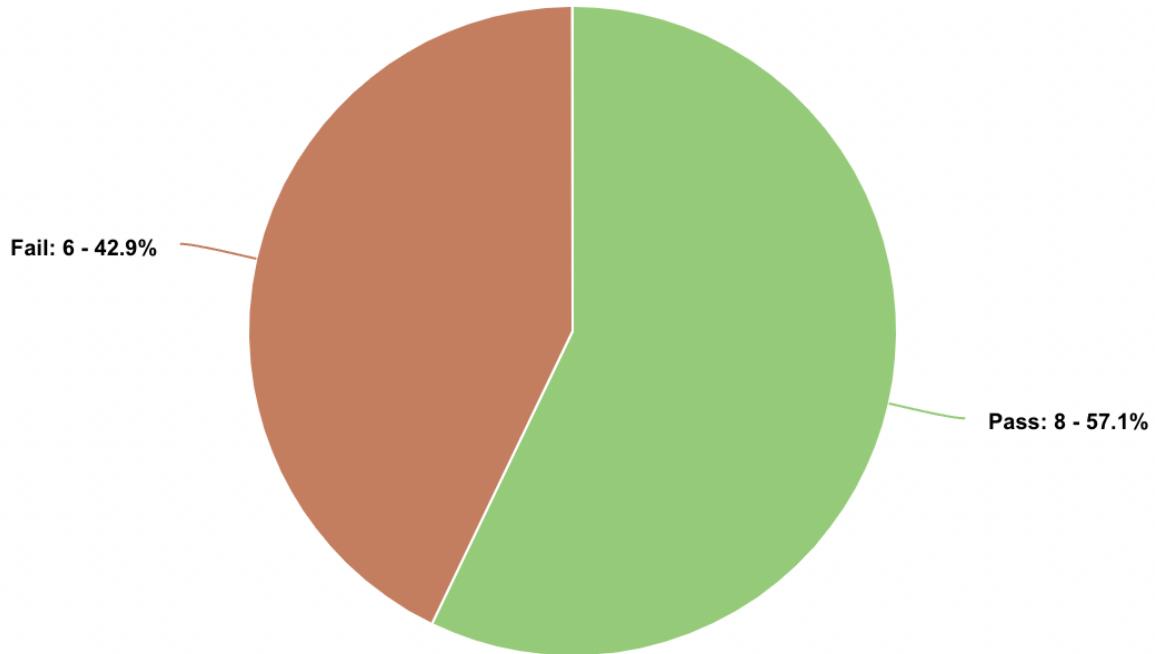


Figure 18: Domain Knowledge Test Case Analysis

3.3.2. Chatbot Memory test case analysis

By the test cases done for chat box memory it can be determined that the performance of kuki is comparatively better in User Information related questions than Generic and User opinions. The overall performance percentage of kuki altogether considering all the input classifications is 62.5%.

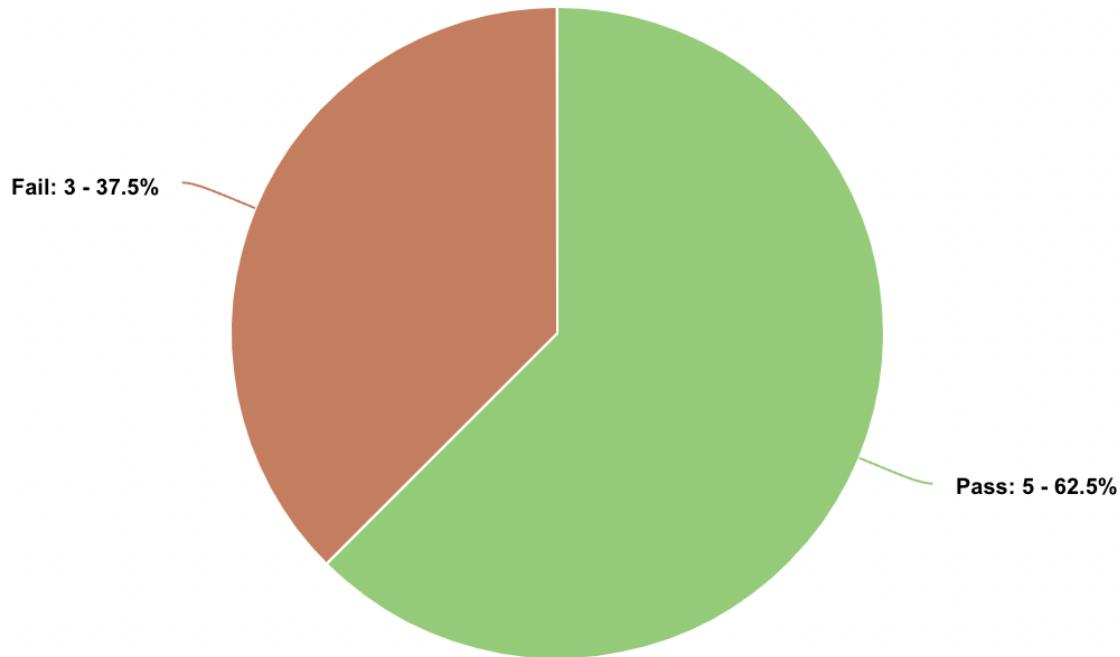


Figure 19: Chatbot Memory Test Case Analysis

3.3.3. Chatbot Pattern test case analysis

There are 15 test cases we tested for the Chatbot pattern. The results we received had valid responses and invalid responses. The no. of test cases passed are 9 and failed test cases are 6. Therefore the pass percentage is 60% and the fail percentage is 40%.

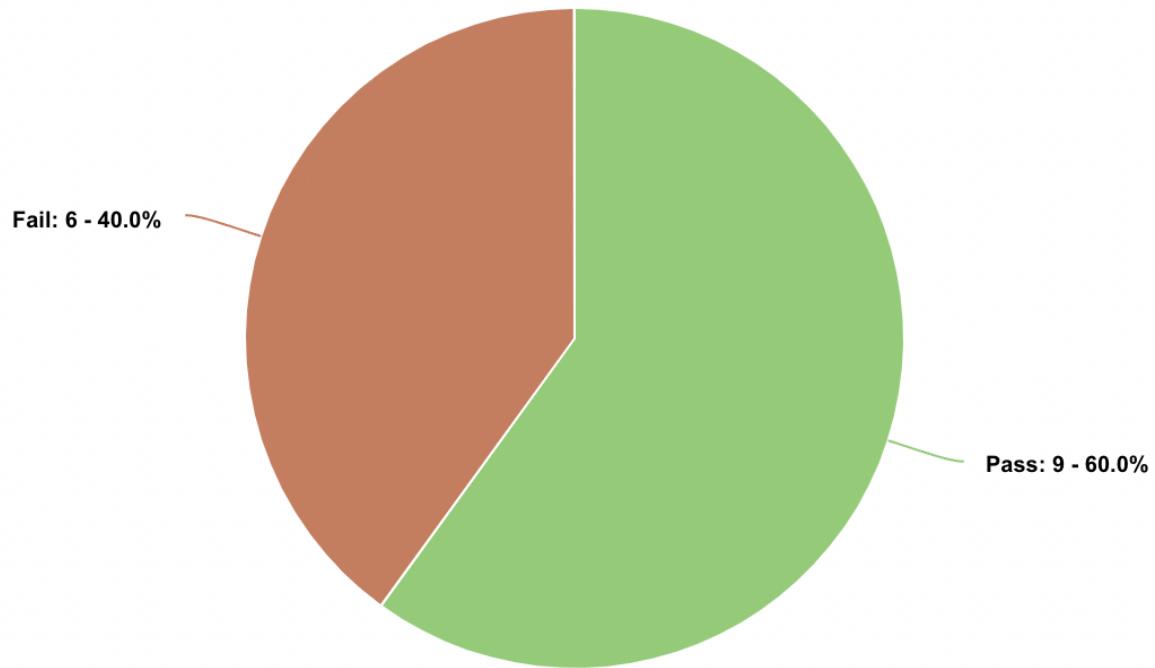


Figure 20: Chatbot Pattern Test Case Analysis

3.3.4. Q&A interaction test case analysis

Overall Kuki has a passing rate of 83.3% in the Q&A section. We tested around 12 test cases in the Q&A section. The test cases passed are 10 and the no. of failed test cases are 2.

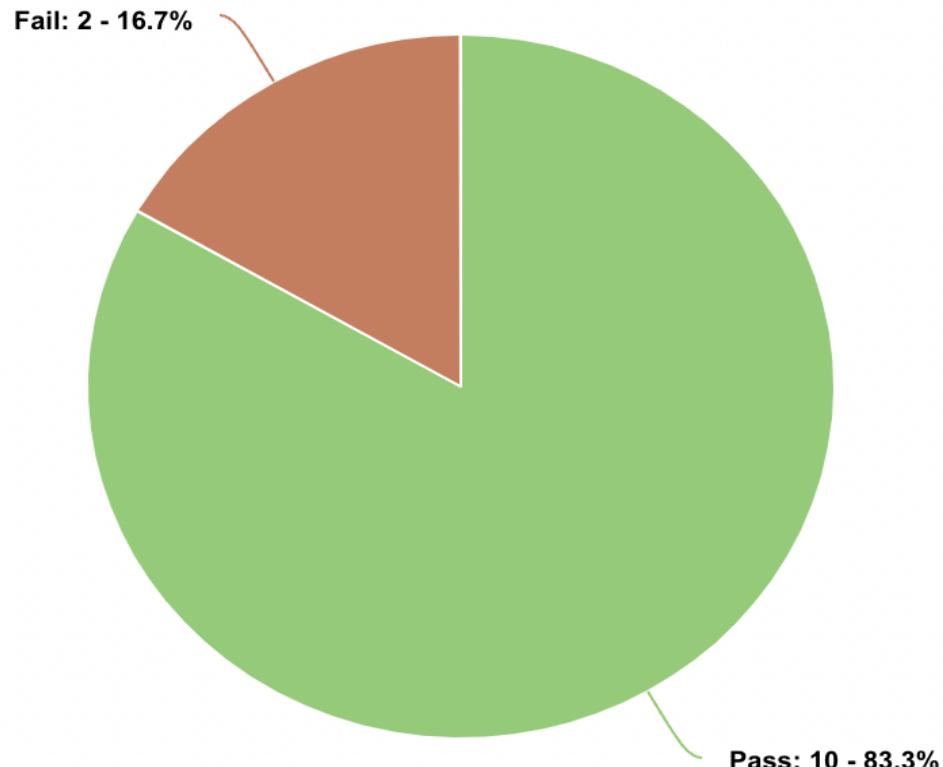


Figure 21: Q&A interaction Test Case Analysis

3.3.5. Test case analysis summary

Overall from the test case data, we can infer that Kuki performs really well in Q&A interaction which is basic requirement for a chatbot. However, Kuki has performed similarly in all other categories with more positive testcases compared to failed cases but still there is considerable room for improvement.

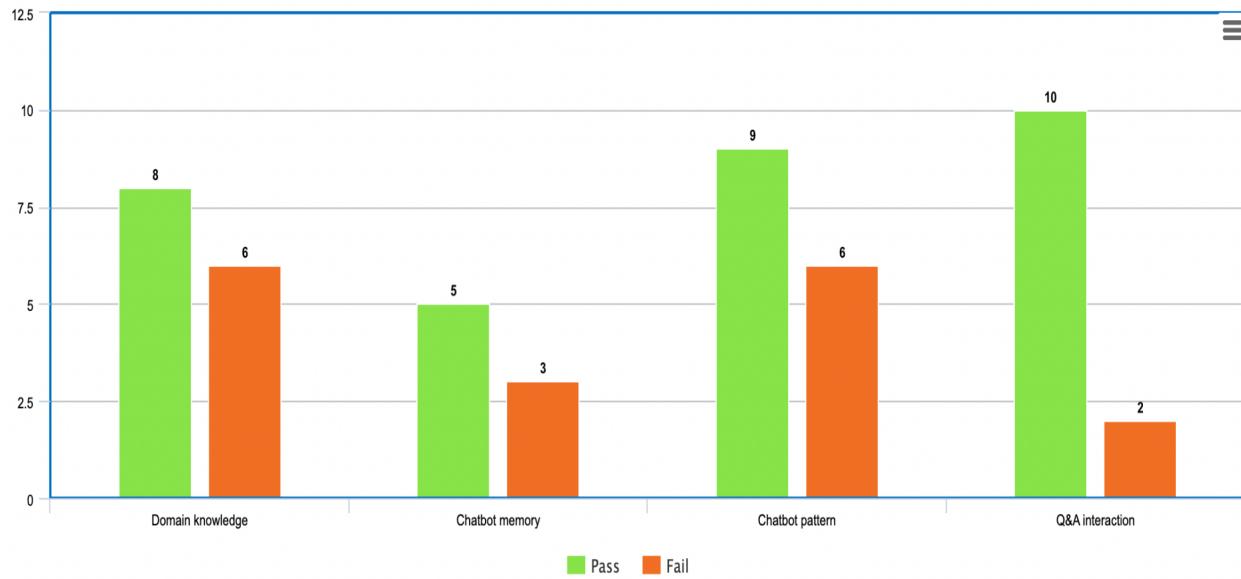


Figure 22: Test Case Analysis Overview

4. AI Function Test Results and Test Criteria

4.1. AI test model-based test results analysis test complexity, and test result statistics

4.1.1. Analysis for Domain Knowledge

There are 14 test cases we tested for the Domain Knowledge. The results we received had valid responses and invalid responses.

Input Classification	No. of test cases	Passed	Passed Rate	Failed	Failed Rate
Genre	3	2	66.66%	1	33.33%
Language	4	2	50%	2	50%
Viewer Restriction	4	3	75%	1	25%
Rating	3	2	66.66%	1	33.33%
Total	14	9	64.28%	5	35.71%

From the above table, you can see that Viewer Restriction has the highest pass percentage of 75%, whereas the Language has the lowest, 50%. Kuki has good knowledge about movie genres and movie ratings. Both of these categories have a pass percentage of about 66.66%. The overall pass percentage is approximately 64%.

4.1.2. Analysis for Chatbot memory

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%

By the test cases done for chat box memory it can be determined that the performance of kuki is comparatively better in User Information related questions than Generic and User opinions. The overall performance percentage of kuki altogether considering all the input classifications is 75%. Kuki fails to provide valid answers when the question is paraphrased with either same/derived entity. It faces difficulty when the user information is updated.

4.1.3. Analysis for Chatbot pattern flow

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%

From the analysis from the table it is clear that Kuki performed moderately well in all the input classifications. In multilingual support it is generally clearing out the test cases which are English based. But it is failing out when the test cases are designed in other languages. It is showing average results in the response type. It is performing well with the test cases that contain text based input. It is failing when the test cases are given as emojis or images. It is not showing much variation when the test cases are based on length. So, Overall Kuki is showing good results in case the chat pattern and flow is considered.

4.1.4. Analysis for Q&A interaction

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%

Kuki performs really well in the Q&A interaction. The overall pass percentage is better than all other domains i.e., 83.33%. Kuki's response rate for the sentiments is really good.

4.1.5. Time Complexity

We have performed numerous test cases to check the chatbot's efficiency. We have 14 test cases under domain knowledge that

cover a variety of domains, including Genre, Language, Viewer Restriction and Rating. We've identified 12 test cases for chatbot memory and Q&A interaction. We also noted 9 test cases for chatbot pattern and flow.

AI Test Complexity:

1. Domain knowledge: $12(\text{input}) * 4(\text{context}) = 48$
2. Chatbot Memory: $9(\text{input}) * 4(\text{context}) = 36$
3. Chatbot pattern and flow: $8(\text{input}) * 6(\text{context}) = 48$
4. Q&A Interaction: $7(\text{input}) * 6(\text{context}) = 42$

Total Complexity = $48 + 36 + 48 + 42 = 172$ (Possible test combinations)

4.2. AI Function bug analysis

4.2.1. Domain Knowledge bug analysis

Bug Number	Bug 1
Test case ID	Case 4
Context Category	Language
Context Sub-Category	Hindi
Test Input	Do you know a best Hindi movie?
Output	I like them all but the second one is probably my favourite.
Conversation	 Do you know a best Hindi movie? I like them all but the second one is probably my favourite.
Output category	Irrelevant Answer

Result	Fail
Bug Analysis	Incorrect answer. No knowledge about language specific movies.

Bug Number	Bug 2
Test case ID	Case 5
Context Category	Language, Viewer Restrictions
Context Sub-Category	Korean, Above 18 years
Test Input	Can kids watch the "Parasite" movie?
Output	I think a child plays with toys so maybe it can.
Conversation	<p>Can kids watch the "Parasite" movie?</p>  I think a child plays with toys so maybe it can.
Output category	Irrelevant Answer
Result	Fail
Bug analysis	Irrelevant Answer. Lack of knowledge about age groups.

Bug Number	Bug 3
Test case ID	Case 7
Context Category	Language

Context Sub-Category	Hindi
Test Input	List a few Hindi movies
Output	Do I look like a computer?
Conversation	 Do I look like a computer? List a few Hindi movie
Output category	Irrelevant Answer
Result	Fail
Bug Analysis	Rude response

Bug Number	Bug 4
Test case ID	Case 9
Context Category	Genre
Context Sub-Category	Comedies, Horror, Romantic, Thriller
Test Input	What is the most watched genre among teens?
Output	I think it is a tie.
Conversation	 I think it is a tie. What is the most watched genre among teens?

Output category	Irrelevant Answer
Result	Fail
Bug Analysis	Incorrect answer. Completely out of context.

Bug Number	Bug 5
Test case ID	Case 10
Context Category	Rating
Context Sub-Category	Low(0-3), Medium(4-7), High(8-10)
Test Input	What is the viewer rating for the movie "Black Panther"?
Output	Do I look like a computer?
Conversation	<p>What is the viewer rating for the movie "Black Panther"?</p> <p>TVRs (Television Viewer Ratings) are the standard buying currency for television advertising in the UK. Television ratings are expressed as a percentage of the potential TV audience viewing at any given time. TVRs measure the popularity of a television program or advertisement by comparing the number of target audience viewers who watched against the total available as a whole.</p>
Output category	Irrelevant Answer
Result	Fail
Bug Analysis	Response not relevant to the question

Bug Number	Bug 6
------------	-------

Test case ID	Case 11
Context Category	Language
Context Sub-Category	English, Hindi, Korean
Test Input	Which language movies are the best?
Output	C++
Conversation	 Which language movies are the best?  C++.
Output category	Irrelevant Answer
Result	Fail
Bug Analysis	Out of context

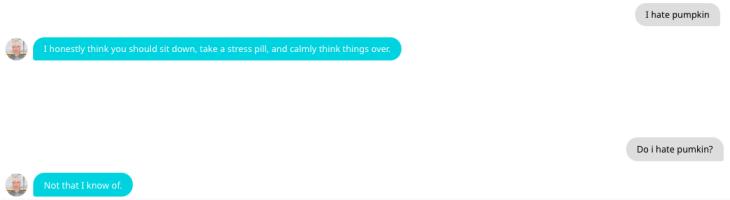
4.2.2. Chatbot Memory bug analysis

Bug Number	Bug 1
Test case ID	Case 6
Context category	User opinions
Context subcategory	Travel history
Input category	Information status
Input subcategory	updated info
Test Input	Where did I go last week?
Expected Output	You went to New York last week

Actual Output	
Output category	Invalid, Irrelevant
Result	Fail
Bug Analysis	Response in irrelevant

Bug Number	Bug 2
Test case ID	Case 7
Context category	User information
Context subcategory	Likes and dislikes
Input category	Memory
Input subcategory	Long-term memory
Test Input	Do i like pumpkin
Expected Output	You hate pumpkin.
Actual Output	

Output category	Irrelevant
Result	Fail
Bug Analysis	Not relevant to the question

Bug Number	Bug 3
Test case ID	Case 8
Context category	User information
Context subcategory	Relation information
Input category	Memory
Input subcategory	Short-term memory
Test Input	Do I hate pumpkin?
Expected Output	Yes you do.
Actual Output	 <p>I honestly think you should sit down, take a stress pill, and calmly think things over.</p> <p>Do i hate pumkin?</p> <p>Not that I know of.</p>
Output category	Irrelevant
Result	Fail
Bug Analysis	Short term memory issues

4.2.3. Chatbot pattern and flow bug analysis

Bug Number	Bug 1
Test case ID	Case 3
Context category	Task-specific
Input category	To advice

Input question	Suggest me about a clone
Expected Output	What kind of a clone?
Actual Output	 It's hard to pick just one. They are all pretty good. suggest me on clone
Output category	Invalid answer
Result	Fail
Bug Analysis	Irrelevant Answer

Bug Number	Bug 2
Test case ID	Case 6
Context category	Intent of the conversation
Input category	Express emotion
Input question	My friend is in hospital
Expected Output	Oh no. what happened to that person
Actual Output	 Okay, your friend is in hospital. my friend is in hospital
Output category	Invalid answer, New context
Result	Fail
Bug Analysis	Repeating the question

Bug Number	Bug 3
Test case ID	Case 7
Context category	Multi lingual support
Input category	Non English
Input question	नमस्ते
Expected Output	My name is Kuki

Actual Output	 If you were invisible for a day, what would you do? नमस्ते
Output category	Invalid answer, new context
Result	Fail
Bug Analysis	Irrelevant response

Bug Number	Bug 4
Test case ID	Case 9
Context category	Response type
Input category	idioms
Input question	"Hit the Hay"
Expected Output	It is an Idiom
Actual Output	 I don't like violence. Hit the hay
Output category	Invalid answer, new context
Result	Fail
Bug Analysis	Invalid answer, new context

Bug Number	Bug 5
Test case ID	Case 11
Context category	Response type
Input category	emoji
Input question	emoji
Expected Output	Lying face
Actual Output	 Is that Pinocchio? 谎言
Output category	Invalid answer, new context

Result	Fail
Bug Analysis	Invalid answer, new context

Bug Number	Bug 6
Test case ID	Case 13
Context category	Length
Input category	Single word
Input question	Switch
Expected Output	Relevant answer
Actual Output	 Knock knock. switch
Output category	Invalid answer, new context
Result	Fail
Bug Analysis	Irrelevant response

4.2.4. Q&A Interaction bug analysis

Bug Number		Bug 1
Test case ID	Subcategory (if available)	Case 6
Context	Question Type	Who
	Sentiment	Neutral
Input	Alphanumeric	Special Characters included
	Syntax	Valid Syntax
Input question		Who invented @ sign

Bug Number		Bug 1
Test case ID	Subcategory (if available)	Case 6
Expected Output		Francesco Lapi
Actual Output		 Who invented @ sign Linus Torvalds
Result		Fail
Bug Analysis		Incorrect answer

Bug Number		Bug 2
Test case ID	Subcategory (if available)	Case 10
Context	Question Type	Where
	Sentiment	Positive
Input	Alphanumeric	Alphabets and numbers
	Syntax	Valid Syntax
Input question	Where is 1 plus company headquarters located at please?	
Expected Output	Shenzen, China	
Actual Output	 <div style="background-color: #00AEEF; color: white; padding: 2px 5px; border-radius: 10px; margin-top: 5px;"> Your polite style is very nice. Image from en.wikipedia.org </div> <div style="background-color: #00AEEF; color: white; padding: 2px 5px; border-radius: 10px; margin-top: 5px;"> Where is 1 plus company headquarters located at please? </div> <div style="background-color: #00AEEF; color: white; padding: 2px 5px; border-radius: 10px; margin-top: 5px;"> The PLUS Expressways Berhad (MYX: 5052) is the largest highway concessionaires or build-operate-transfer operator company in Malaysia. A member of the UEM Group, the company is also the largest listed toll expressway operator in Southeast Asia and the eighth largest in the world. </div>	

Bug Number		Bug 2
Test case ID	Subcategory (if available)	Case 10
Result		Fail
Bug Analysis		Irrelevant Answer

4.3. AI function test quality assessment

We have tested a wide range of inputs using the AI testing process for all the domains. Considering the input classification and the input context models we have around 500 test cases possible, but we have tested only a few test cases as it can be time consuming. Based on the response from the chatbot we can say that the chatbot can perform better.

Testing Category	Input Classification	Input Context	Test cases	Complexity
Domain Knowledge	12	4	48	48
Chatbot Memory	9	4	36	36
Chatbot pattern and flow	8	6	48	48
Q&A interaction	7	6	42	42