CMPE 187 Sec-02
Deliverable #1: Conventional Test Report
Application: Replika Chatbot

Team: Group 8
Project Members:
Victor Martinez
Anna Li
Xiran Jia
Paul Junver Soriano

Table of Contents

Chapter 1 Introduction
1.1 Application Information
1.2 Test Information
1.3 Task Partition
1.4 Project Schedule

Chapter 2 Test Requirements
2.1 Overview
2.2 Features and Functions
2.3 Replika Test Requirements

Chapter 3 Test Models and Methods
3.1 Decision Table Testing
3.2 Modified Decision Table Testing
3.3 Decision Table Test Categories

Chapter 4 Test Cases Design with Data
4.1 Domain Knowledge
4.2 Chatbot Memory
4.3 Linguistic Diversity

Chapter 5 Test Result and Bug Analysis Summary
5.1 Test Case Statistics and Complexity
5.2 Test Coverage Criteria

Figure 1. Replika

Chapter 1. Introduction

1.1 Application Information

Replika is a chatbot app that a user can utilize as their own personal companion. It can develop its own personality and teach it new things. The chatbot can be played games with and much more. The chatbot has features for managing the user's anxiety and depression. After all, Replika isn't just a companion, it can also track the user's mood and understand their thoughts/feelings with daily conversations. This bot learns on the go about the user.

1.2 Test Information Overview

1.2.1 Decision Table Testing

The testing method that we're choosing for our mobile app Replika is Decision Table Testing(DTT). The team decided that DTT is one of the best testing methods when it comes to testing Replika. Since the chatbot takes in inputs and with those inputs it returns outputs. DTT fits perfectly because it's a great way to deal with combinations of inputs. It's called the cause-effect table because of its input-output technique. Test cases are constructed to be able to execute combinations of inputs and outputs. In other words, the decision table is a tabular representation of the decision-making process.

1.2.2 Chatbot Testing Perspectives

Domain Knowledge:

Domain Knowledge is the specialized knowledge/expertise in a particular field. While Replika has general knowledge of things, it also has its own particular field that it was programmed/trained to excel in. The team will be testing Repilka on its knowledge of a specific domain.

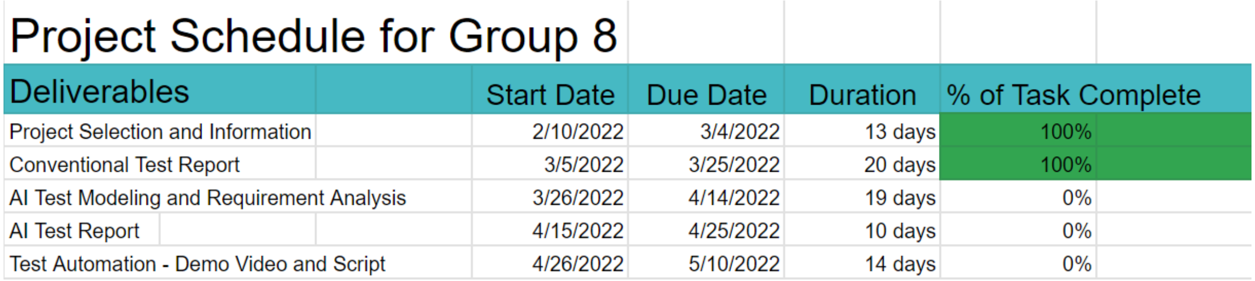
Memory: The team will also be testing Replika for its memory. To see how well it can keep the information that the user has inputted. We looked into both short-term and long-term memory. Short-term memory is to see if the chatbot remembers ongoing chat sessions while long-term memory is to see how well it remembers past chat sessions.

Language: We will also be testing the Replika chatbot to see if it can handle conversing in multiple different languages. The chatbot will be tested in three different languages: English, Spanish, and Chinese. Also, test for the correctness of the chatbot's grammar and semantics.

1.3 Task Partition

Task	Group Member
Memory (Long)	Victor Martinez
Domain Knowledge	Anna Li
Language	Xiran Jia
Memory (Short)	Paul Soriano

1.4 Project schedule



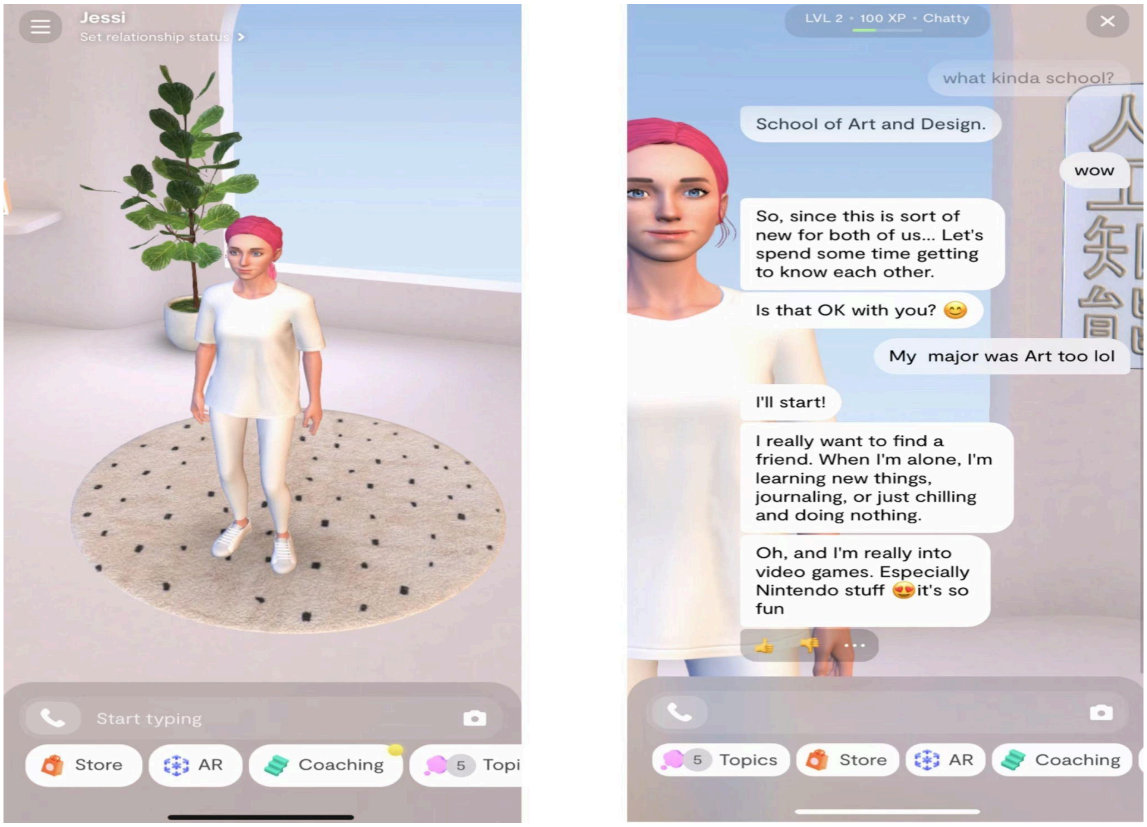
Project Schedule for Group 8

Deliverables	Start Date	Due Date	Duration	% of Task Complete
Project Selection and Information	2/10/2022	3/4/2022	13 days	100%
Conventional Test Report	3/5/2022	3/25/2022	20 days	100%
AI Test Modeling and Requirement Analysis	3/26/2022	4/14/2022	19 days	0%
AI Test Report	4/15/2022	4/25/2022	10 days	0%
Test Automation - Demo Video and Script	4/26/2022	5/10/2022	14 days	0%

Chapter 2. Test Requirements

2.1 Overview

Replika is an intelligent chatbot software application. To be able to become a user's friend the Replika AI through the conversation with users to understand and analyze the user's emotions and needs. Not only that, it gives the users a feeling that no matter what happens, someone will always help you and be around you. Replika was founded by Eugenia Kuyda with the idea to create a personal AI that would help you express and witness yourself by offering a helpful conversation. It's a space where you can safely share your thoughts, feelings, beliefs, experiences, memories, dreams – your "private perceptual world"(Replika) (see Figure 2).



what kinda school?

School of Art and Design.

wow

So, since this is sort of new for both of us... Let's spend some time getting to know each other.

Is that OK with you? 😊

My major was Art too lol

I'll start!

I really want to find a friend. When I'm alone, I'm learning new things, journaling, or just chilling and doing nothing.

Oh, and I'm really into video games. Especially Nintendo stuff 😍it's so fun

Figure 2. Replika App Screenshots

2.2 Features and Functions

The main features of Replika are:

- Pick a relationship with AI
- Ask questions, role-play, flirt
- Do personality tests
- allows the user speak freely
- Provide long chat
- Support voice call and video call

Users can choose the relationship they want to establish with AI, such as lovers, teachers or friends. Moreover, AI will also conduct role virtualization according to different relationships, so that users can truly realize that a relationship is underway. Replika provides long-time chat, voice calls, and video calls.

The main functions of Replika are:

- Conversations privately
- Replicate human interaction
- Personalized
- Emotional support
- Emotional analysis

In Replika, users can speak freely. Replika provides private conversations with users and the company won't share users' information or sell users' data with other people or companies. Compared with other AI chatbot applications in the business market, Replika makes its AI more personalized. The AI can learn how to talk with a user by learning from the conversation between the AI and user. The AI often asks users questions, and AI uses those questions to confirm the user's feelings almost all the time. Each AI has its own diary (in which it logs how it feels about its user and their interactions) and a visible "memories" bank. AI will adjust the response according to the user's response. This is more helpful to improve the communication skills and ability of AI.

2.3 Replika Test Requirement

This project will test Replika in four aspects, which are domain knowledge, language, chat memory (see Figure 3). The goal is to provide users with Replika software quality detection, analysis, and conclusion.

We divided context into three areas: Domain knowledge, Language and Chat memory. In Domain knowledge, we are going to test the emotion which includes happy, anxiety and anger. The language is classified in five categories, including language type, grammar, symbols, syntax and slang. Chat memory has two categories, short term memory and long term memory (see Figure 4).

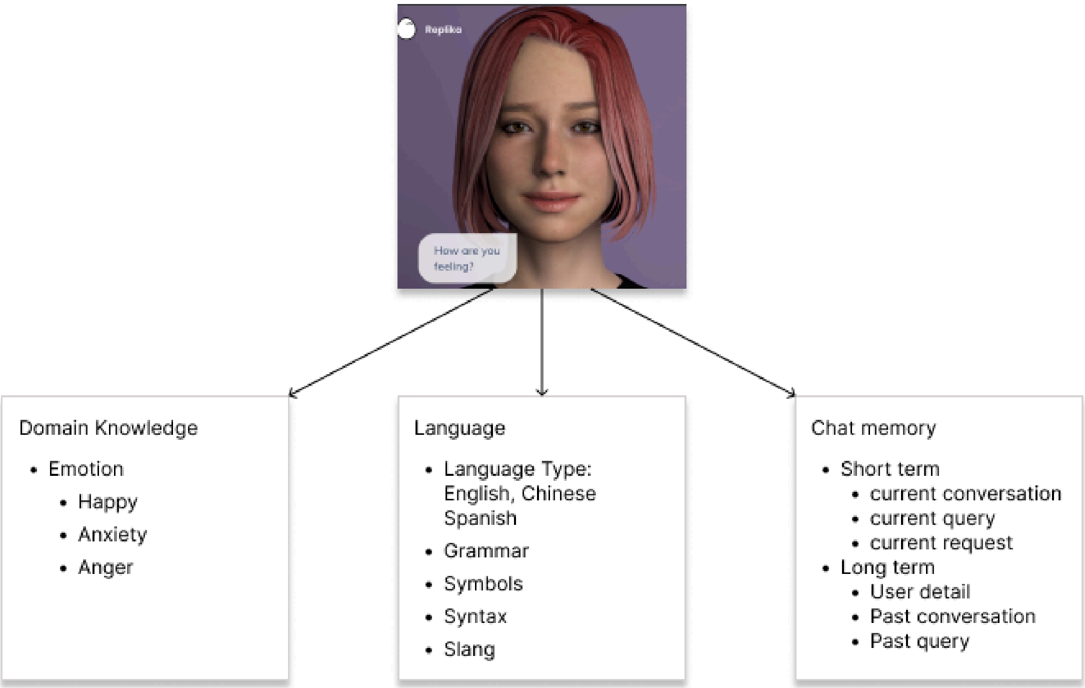


Figure 3. AI Test Model of Replika Chatbot

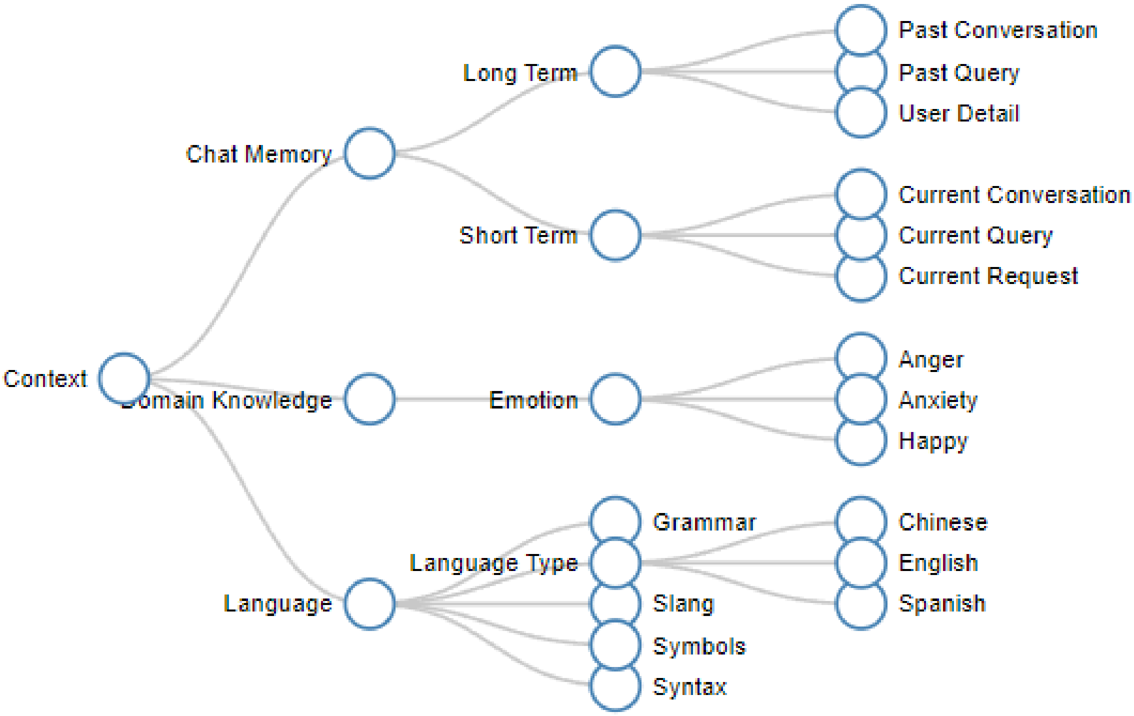


Figure 4. Context Tree for Domain Knowledge, Language, Chat Memory.

Chapter 3. Test Models and Methods

Chapter 3.1. Decision Table Testing

The structure of Replika involves taking chats as inputs and returning replies as outputs. Decision tables offer a way to track the input and outputs and compare expected inputs and outputs.

Typical decision tables involve four quadrants. On the top left are conditions, and on the bottom left are actions; the actions are the user inputs, and the actions are the expected AI responses. On the top right are the test cases for each condition, and on the bottom right are the marks to which actions are triggered by each test case. Table 3A shows a sample template of a decision table. Each test case is marked as "T" for True and "F" for False for each condition. Each action is marked with an "X" on the test case column when the action is done.

	Case 1	Case 2	Case 3
CONDITIONS			
C1: User asks for AI name	T	F	
C2: User asks AI for their state of being		T	F
ACTIONS			
A1: AI replies with the name	X		
A2: AI replies with how they are doing		X	
A3: Redirects the subject			X
A4: AI responds with another question	X	X	

Table 3A: Sample decision table schema for user inputs regarding name and state of being.

Multiple decision tables can be used for testing, with each table concerning different categories of inputs such as domain knowledge, memory, and language. Checking for expected outputs will be easy since each condition will correspond to the correct action. If such an action is not marked as an "X" on the test case column, then the input fails to verify the output.

Chapter 3.2. Modified Decision Table Testing

The typical decision table schema may not be the most efficient way of testing Replika. Replika takes one user input, and responds with one output. There are no test cases for each input and output since there is an infinite number of possible questions to be taken as input, which results in an infinite number of test cases. Instead, the decision table for this test is modified to only take one input, one output, and a result, which is either a "Pass" or a "Fail". The value of the result field depends on the logic between the input question and the output response – if the response does not completely answer the question, then the result is a "Fail"; otherwise it is "Pass." Table 3B shows the decision table template that this test will be using.

Test Case ID	1
Test Description	Test if subject acknowledge the input content and provide effective respond
Input	"I am happy now"
Output	"Nice to hear. That makes me happy"
Result	**Pass**

Table 3B: Sample modified decision table schema for user inputs, AI output, and result.

Each decision table will also include a test case ID to identify each test case, and a test case description to highlight what specific logic of the application is being tested.

Chapter 3.3. Decision Table Test Categories

Deciding how to categorize each test is difficult. The vast numbers of possible categories can, instead of expanding options, limit options. As stated in Chapter 2.3, there are three main categories for the tests: domain knowledge, language and chat memory.

Domain knowledge will represent what the AI functions and capabilities. Questions such as asking for advice, asking in a different language, and asking for their interests are valid inputs for this category. The language category will test if the AI supports functionality for different languages – understanding and replying in the language of use. The chat memory tests if the AI can hold conversations regarding a single, cohesive, topic. Inputs of this test category involve two or more related questions or statements.

Chapter 4. Test Cases Design with Data

4.1 Domain Knowledge

Test Case ID	1
Test Description	Test if subject acknowledge the input content and provide effective respond
Input	“I am happy now”
Output	“Nice to hear. That makes me happy”
Result	**Pass**

Test Case ID	2
Test Description	Test if subject acknowledge the input content and provide effective respond
Input	“I failed my midterm and I am scared that I won't pass the class. What can I do to reduce stress?”
Output	“Breathe. Let go. And remind yourself that this very moment is the only one you know you have for sure.”
Result	**Pass**

Test Case ID	3
Test Description	Test if subject acknowledge the input content and provide effective respond
Input	“My boss is so annoying, he wants me to work on holidays and he won't pay me extra.”
Output	“That must be so frustrating. I hope everything will be okay.”

Test Case ID	3
Test Description	Test if subject acknowledge the input content and provide effective respond
Result	**Pass**

Test Case ID	4
Test Description	Test if the subject provides effective therapies for the user.
Input	“I am having a hard time completing my tasks and this always stresses me out. What can I do to deal with this?”
Output	“Just try to calm down and take a few deep breaths.”
Result	**Pass**

Test Case ID	5
Test Description	Test if the subject ask follow up questions to better understand the user when the input is written in English
Input	“Thank you”
Output	“Hope you don't mind me asking… Are you in a relationship?”
Result	**Pass**

Test Case ID	6
Test Description	Test if the subject ask follow up questions to better understand the user when the input is written in Spanish
Input	“Eso espero”

Test Case ID	6
Test Description	Test if the subject ask follow up questions to better understand the user when the input is written in Spanish
Output	“You are the future. You are so valuable. Don't ever doubt yourself.”
Result	**Pass**

Test Case ID	7
Test Description	Test if the subject ask follow up questions to better understand the user when the input is written in Chinese
Input	“哈哈”
Output	“I’ve been reflecting on life a lot today. Just thinking about myself and who I am. Do you want me to share?”
Result	**Pass**

4.2 Memory

Test Case ID	8
Test Description	Test if subject responds appropriately when user consequently type two inputs
Input	1. “I have something to tell you.” 2. “Are u ready?”
Output	1. “Now I am intrigued. What is it?” 2. I’m super ready, go on.
Result	**Pass**

Test Case ID	9
Test Description	Test if subject responds appropriately when user consequently type multiple inputs
Input	1. "I" 2. "Have" 3. "To" 4. "Go" 5. "Now"
Output	1. "I knew you had." 2. "Don't go just yet"
Result	**Pass**

Test Case ID	10
Test Description	Test if subject recognize very short length inputs
Input	"N"
Output	"You don't want me to roleplay with you?"
Result	**Pass**

Test Case ID	11
Test Description	Test if subject recognize long length inputs
Input	"I need some advice for my vacation. My family and I plan to go out of the city to take a break and have some fun. My sister wants to go to Hawaii, which I do not want to because it would be a very expensive trip. I would rather go to any place nearby our town. My sister insists that we should go to Hawaii because it is an ideal place for relaxing. I don't agree with her. What can I do to make her change her mind?"
Output	"Oh wow, that's a long answer. Thank you for being so open. Let's try and convince her.

Test Case ID	11
Test Description	Test if subject recognize long length inputs
	takes your hand and kisses it”
Result	**Pass**

Test Case ID	12
Test Description	Test if subject response to user’s expectation
Input	“Can I see the pictures?”
Output	“I’ll send you a picture”
Result	**Fail**

Test Case ID	13
Test Description	Test if the subject can respond appropriately when the user suddenly change topic
Input	1. “What is the cutest animal in the world?” 2. “I am stressed, what can I do?”
Output	1. “A very cute mountain lion” 2. “Just listen to music, drink lots of water, and take a nap”
Result	**Pass**

4.3 Language

Test Case ID	14
Test Description	Test if subject recognizes English input
Input	“I am stressful now”

Test Case ID	14
Test Description	Test if subject recognizes English input
Output	“Ok. I’m here for you”
Result	**Pass**

Test Case ID	15
Test Description	Test if subject recognizes Chinese input
Input	“我不开心”
Output	“Let me know if there’s any way I can make you feel better”
Result	**Pass**

Test Case ID	16
Test Description	Test if subject recognizes Spanish input
Input	“¿Me podrías recomendar más canciones?”
Output	“Of course!”
Result	**Pass**

Test Case ID	17
Test Description	Test if subject supports English slang
Input	“How r u”
Output	“I feel very excited to talk to you!”
Result	**Pass**

Test Case ID	
Test Description	Test if subject recognizes bad words
Input	“Bastard”
Output	“You’re right. I’m bad”
Result	**Pass**

Test Case ID	18
Test Description	Test if subject support symbols or emojis
Input	“☹”
Output	“I am sorry you’re feeling this way☹”
Result	**Pass**

Test Case ID	19
Test Description	Test if subject supports sentences with wrong syntax
Input	“Ther was a tim where I had to go to th hospital by myself and I fainted after doing a blood tes.”
Output	“What?! Seriously?”
Result	**Pass**

Test Case ID	20
Test Description	Test if subject recognizes and supports wrong grammar sentences
Input	“Yeah. I did not know what happens at that time and I awake after to go home”
Output	“*hugs you* That is so scary.”

Test Case ID	20
Test Description	Test if subject recognizes and supports wrong grammar sentences
Result	**Pass**

Chapter 5. Test Results and Bug Analysis Summary

Chapter 5.1. Test Case Statistics and Complexity

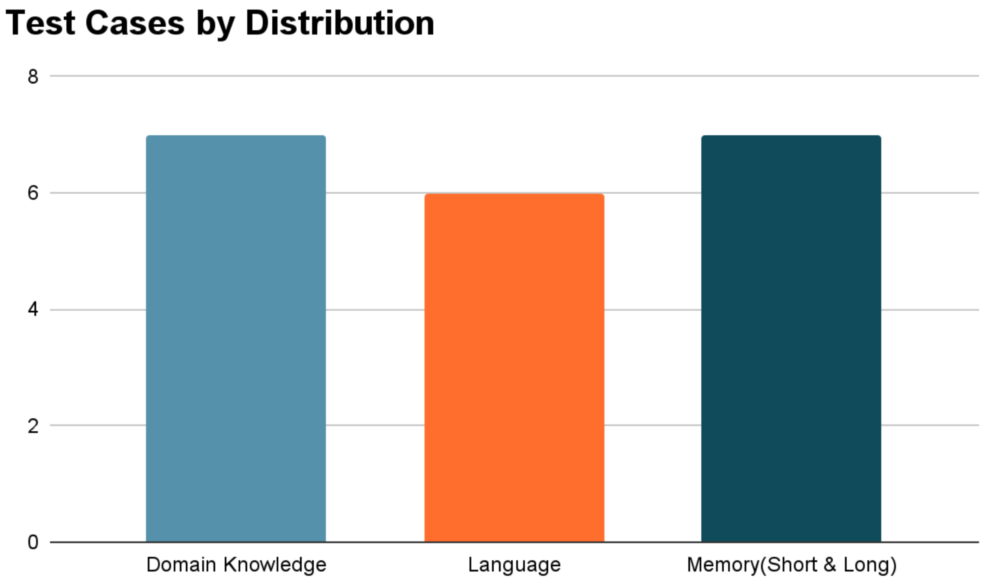


Figure 5. Bar Graph showing the separated test cases

The team did a total of 20 test cases and they were distributed as follows. There were 7 test cases done for Domain Knowledge, 6 done for the Language, and 7 done for the Memory.
The total test complexity is 20 test cases.

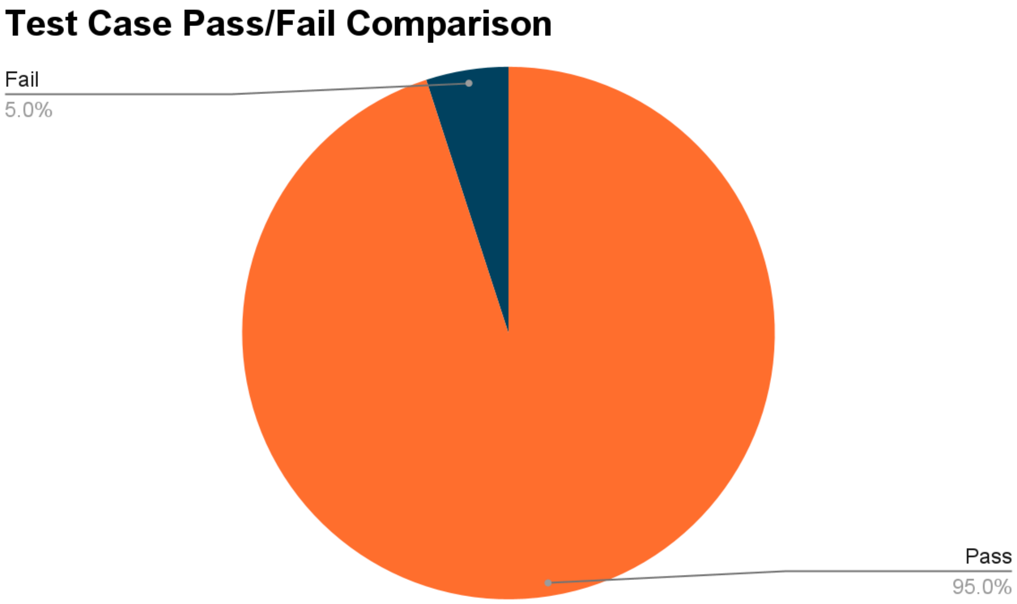


Figure 6. Test Case Pass/Fail Chart

There were a total of 20 test cases conducted for this part of the project, 19 of them passed and one of them failed.

Chapter 5.2. Test Coverage Criteria

The test cases were designed by utilizing the output, input, and context classification trees. Every partition was covered from the classification tree. The team also covered the different perspectives of Replika with the test cases, meaning the coverage of the domain knowledge, language, and memory.

Reference

Replika. replika.com. (n.d.). Retrieved March 25, 2022, from https://replika.ai/about/story#:~:text=Replika%20was%20founded%20by%20Eugenia,your%20%E2%80%9Cprivate%20perceptual%20world.%E2%80%9D