SAN JOSÉ STATE UNIVERSITY

AI Testing – Case Study and Experience

Presented by: Jerry Gao, Professor, and Director

San Jose State University – Excellence Research Center on Smart Technology, Computing, and Complex Systems

Some Mobile AI System Examples

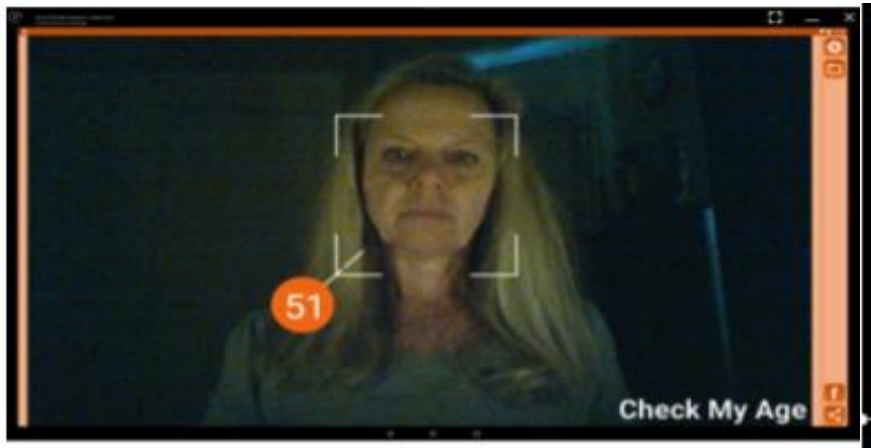
Virtual Personal Assistant	Google’s Allo, Apple’s Siri, Microsoft’s Cortana
Smart Cars	Google’s autopilot project
Music and Video Recommendation systems	YouTube, Netflix
Video games	Call of Duty, Star Craft II, FIFA Journey
Prediction for Retail giants	Amazon, Walmart
Fraud Detection for Banks	Visa, Mastercard, PayPal
Customer Support	Using chatbots to communicate, ex Lenovo.com, Dell.com
Smart Home Devices	Lights turn on as you walk

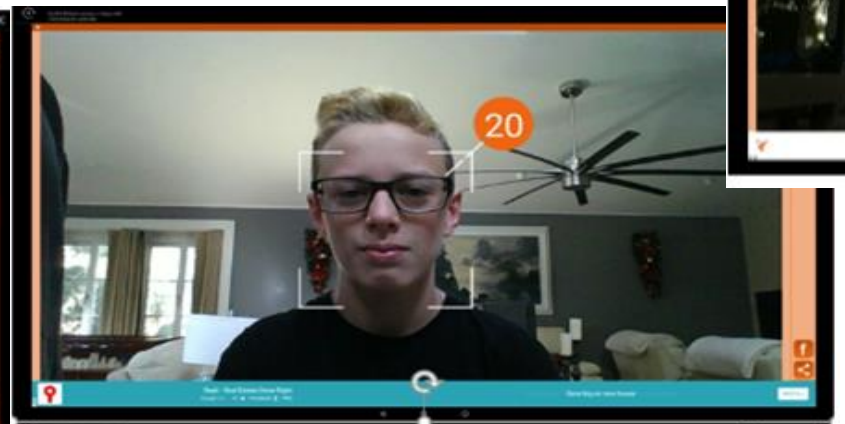
AI Software Testing – Case Study I – Tell Me Age

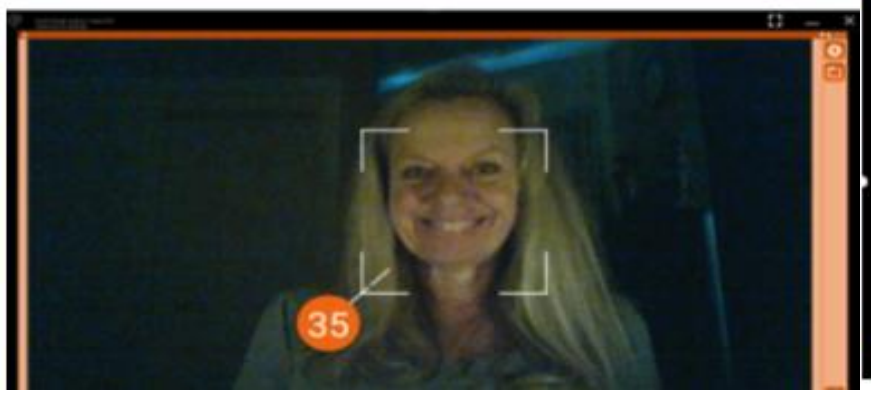
AI System Testing – A Case Study – Tell Me Age

Accuracy is affected by:

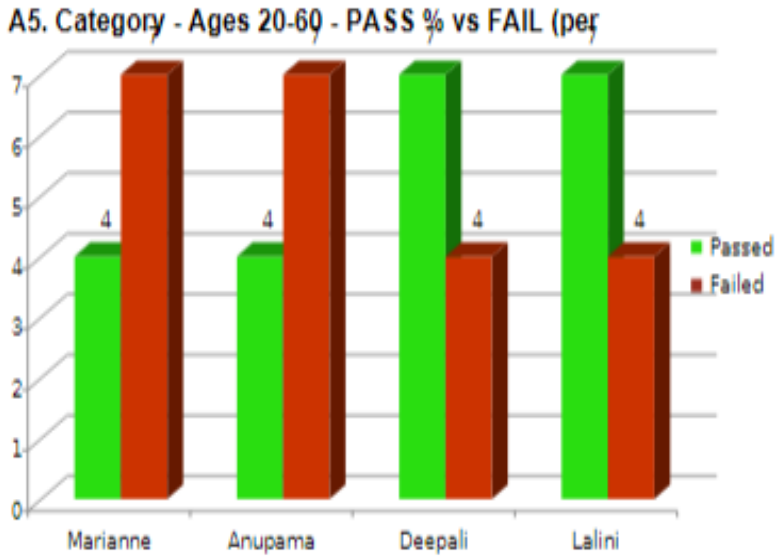
- Lighting Condition
- Personal Ware / Hair Style,…..
- Background Objects in the image

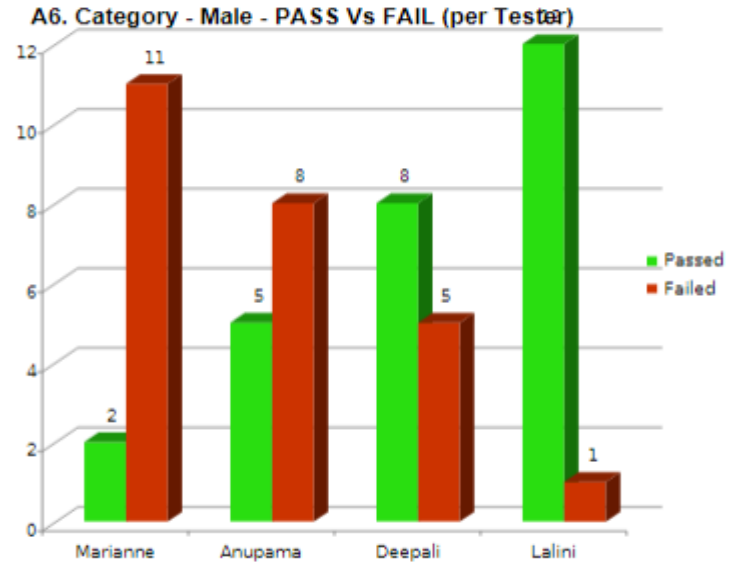


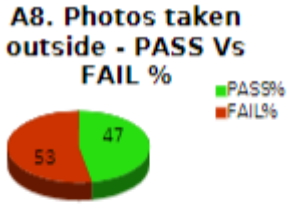


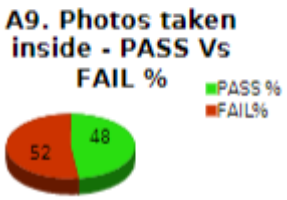


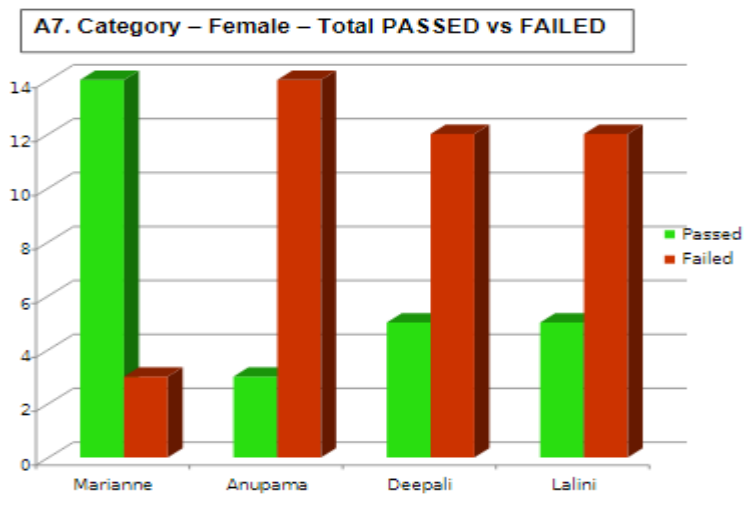
AI System Testing – A Case Study – Tell Me Age



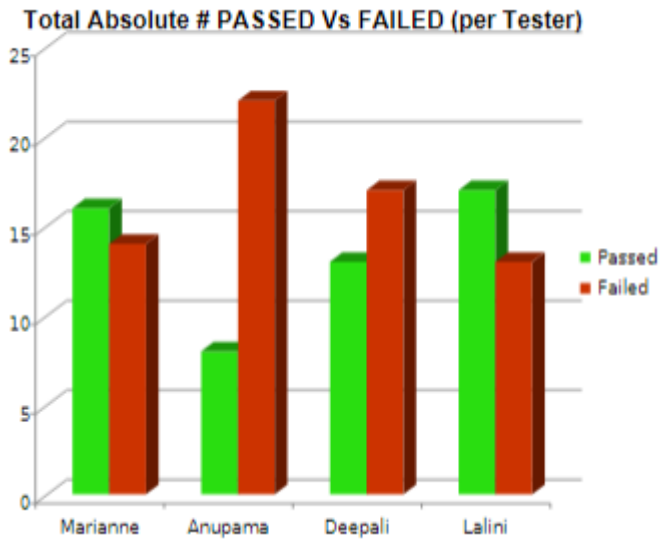


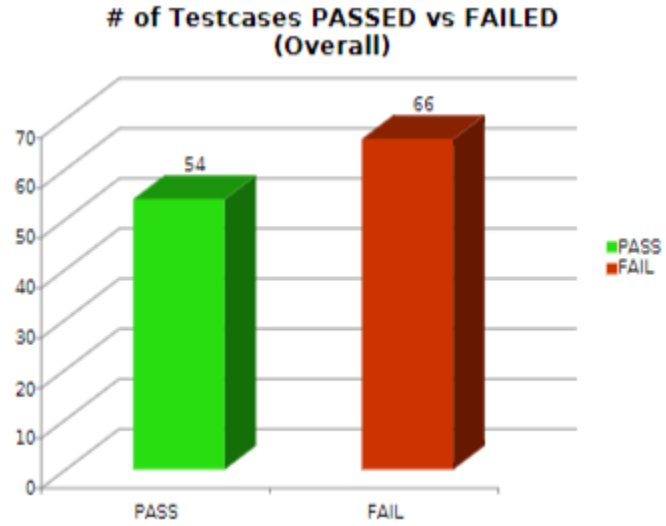


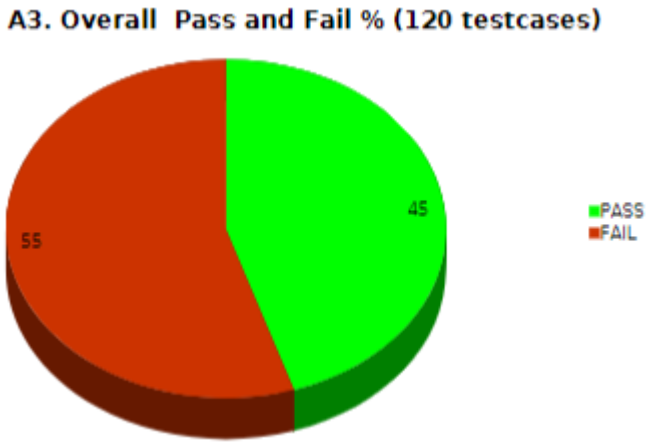


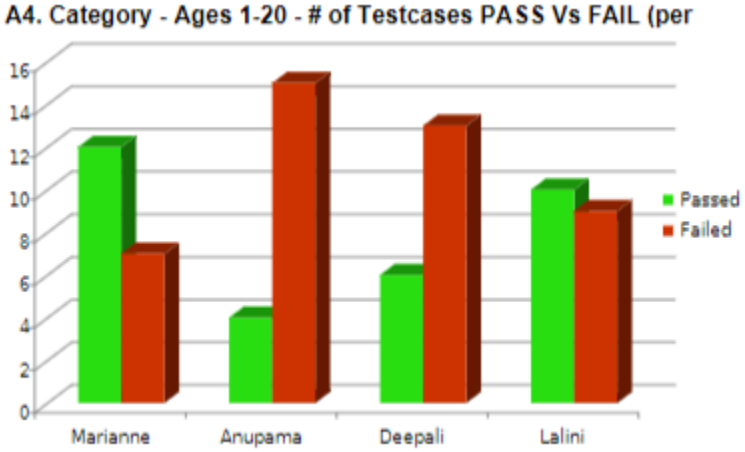


AI System Testing – A Case Study – Tell me Age



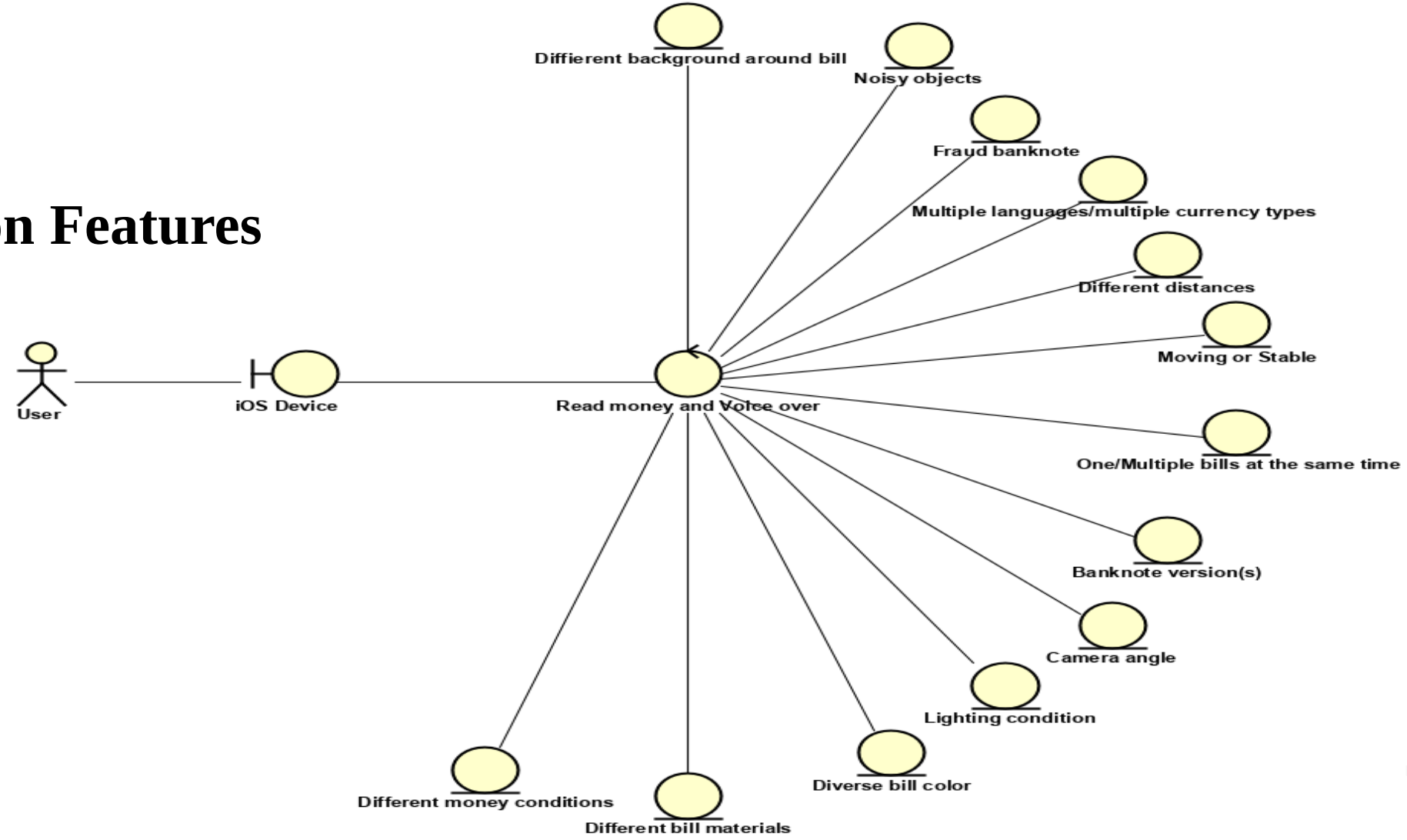






Testing LookTel Money Reader AI Functions

AI Function Features



USING CONVENTIONAL TESTING METHOD

To leverage the testing effectively in a short time, we apply some basic standard conventional testing methods to this mobile app.

- **Decision Table**

  Decision table testing is a kind of testing is more like a cause-effect testing. This testing will determine what kind of output will be obtained on giving various kinds of input with different kind of circumstances. Security holes can be detected in this method. The testing coverage area will be as follow: a) Checking multiple testing condition setup; b) Validate the money reader over some kinds of money material.

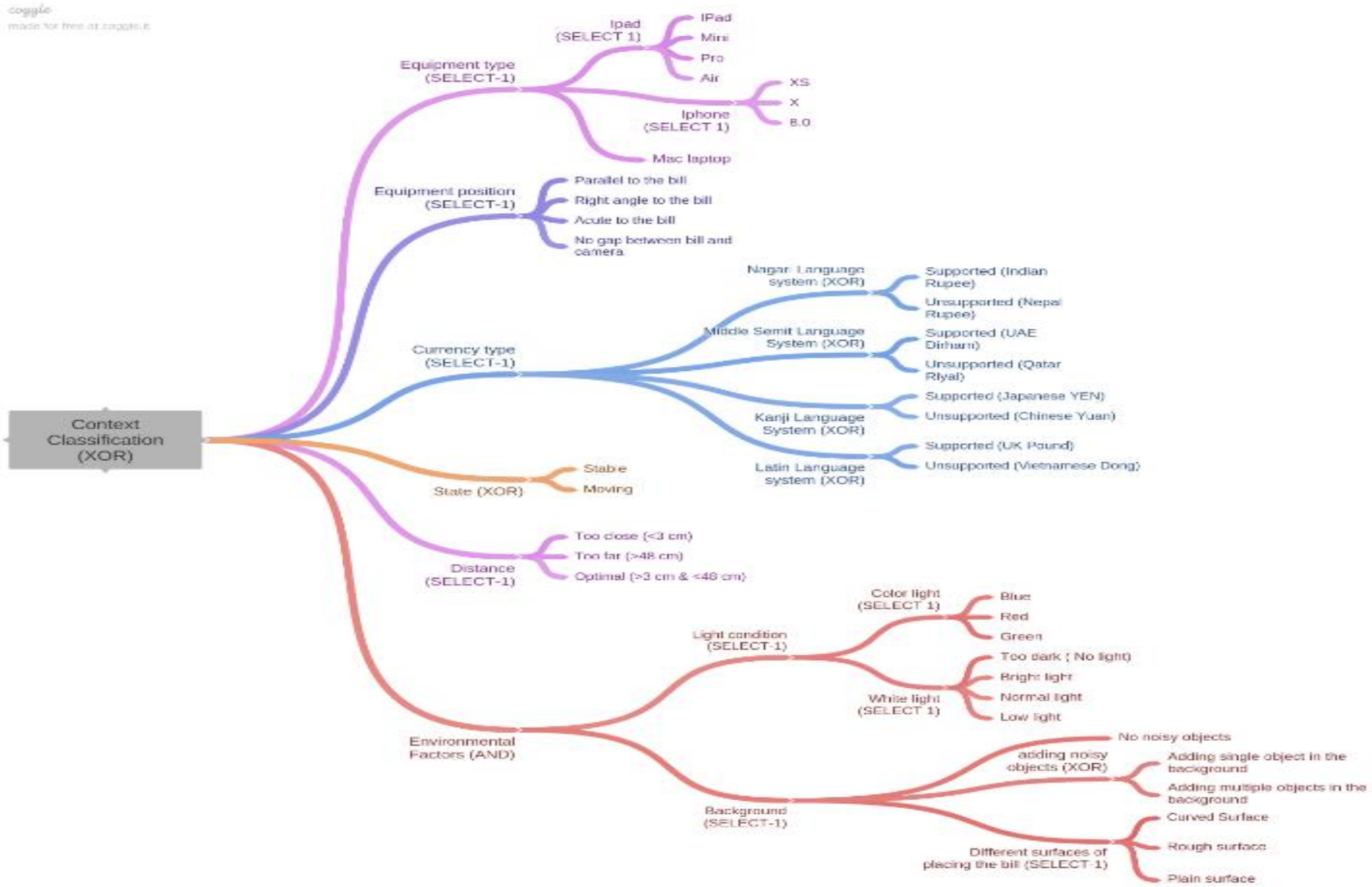
- **Scenario testing**

  This method is helpful to go through from the beginning state to any states of product without any technical view, only focus on user real experience. In this application, some test cases cannot be created by decision table or equivalence partition because it depends on testing setup and supported features on mobile, not application itself. That's the reason why we need to use scenario. Test coverage: This method will be used in testing application on different types of supported equipment, with supported features on each, like: voice over setup, voice quality, text displaying, …

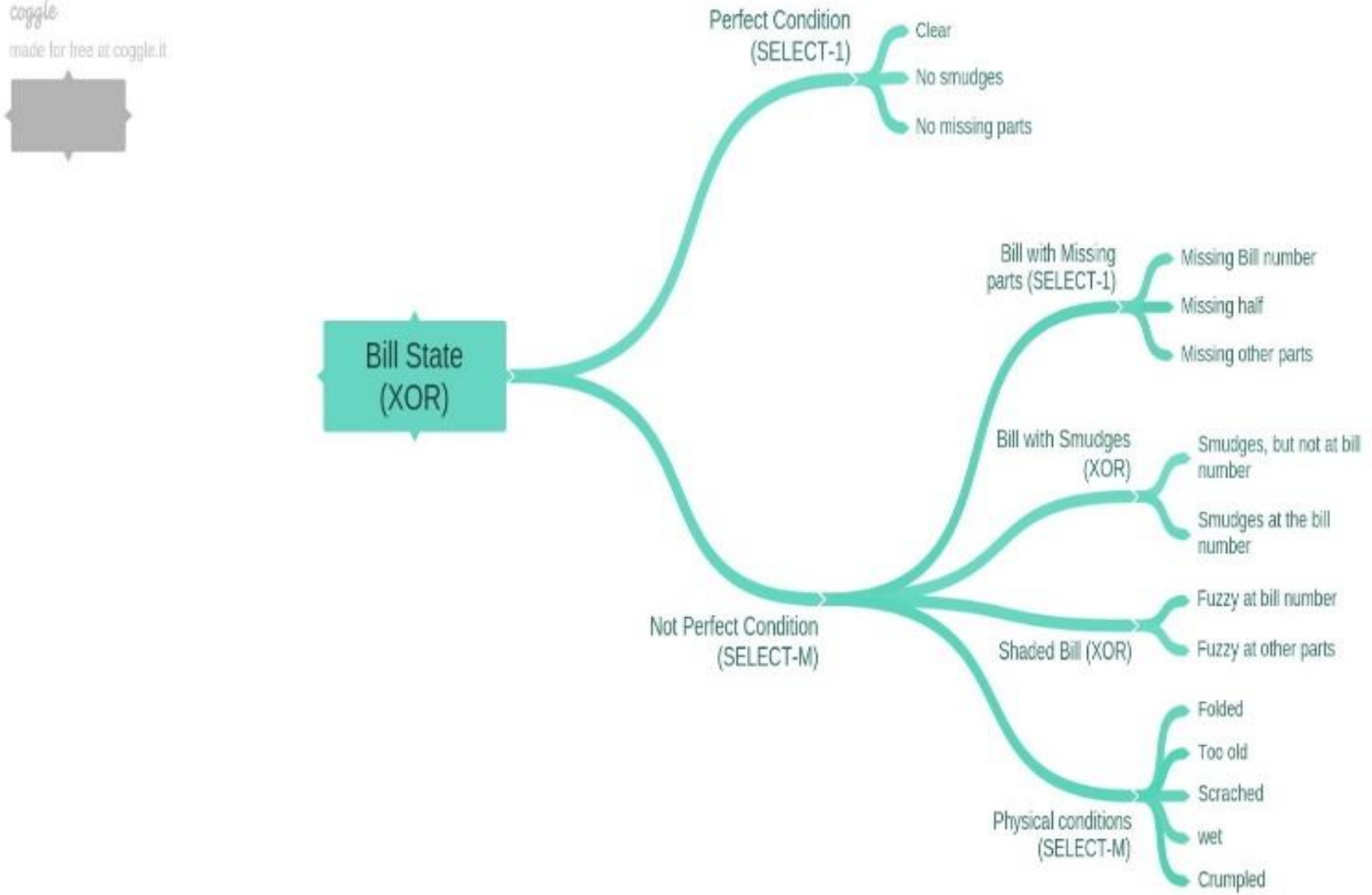
- **Equivalence partitioning**

The application involves in detecting the bills from many countries and each country will have different set of bills of different value. This gives us a wide range of input, further which can be partitioned into classes. From each class, a particular kind of input is chosen as given as input. This kind of testing saves the time of the tester and reduces the tester's workload. Test coverage: This method will help us to cover testing the feature of checking money reader's detection and voice over capabilities on some range of language system around the world. Base on some defined test methods above, we designed our test cases for each business requirement domain follow each testing technique.

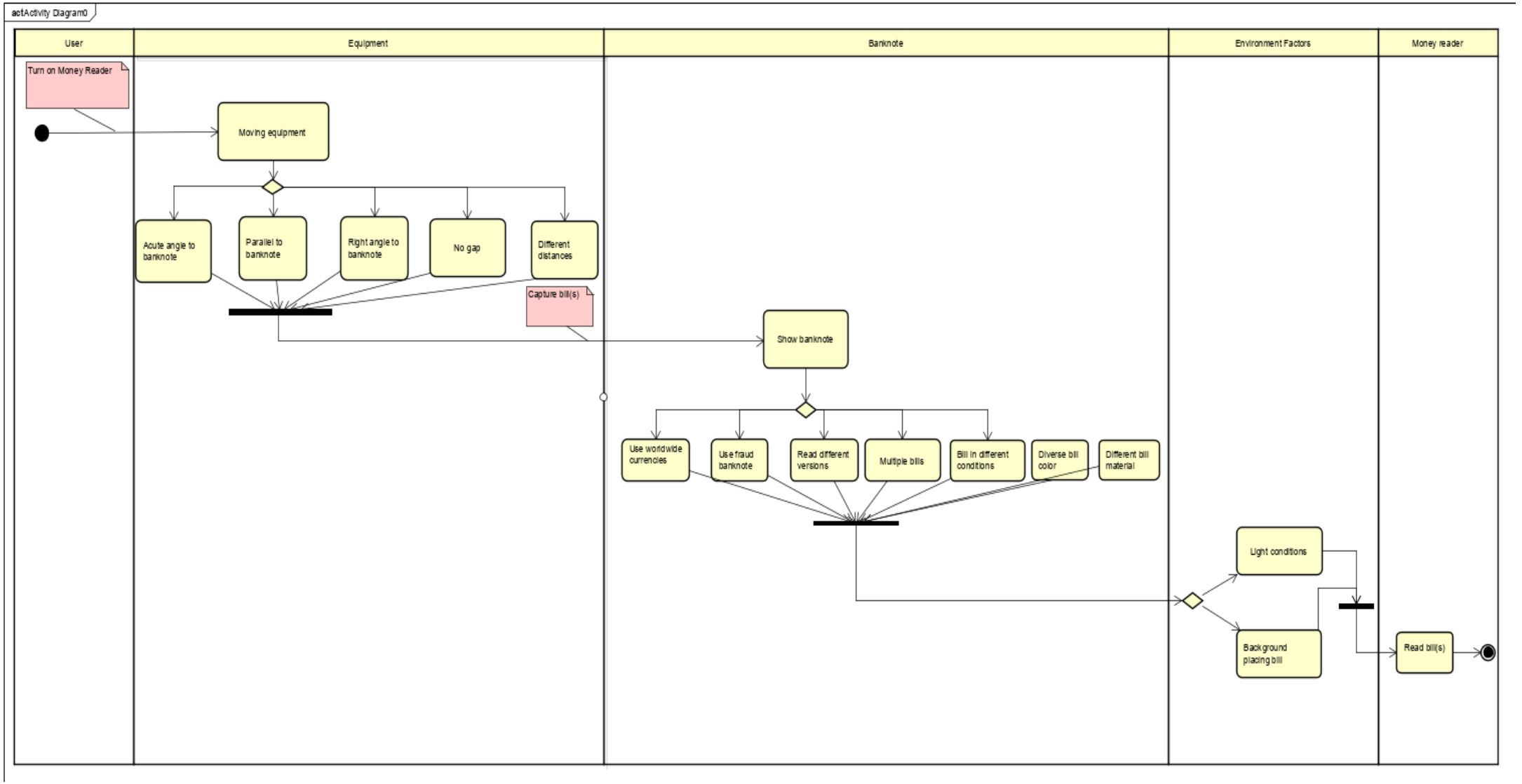
3D Money Reader - Context Tree



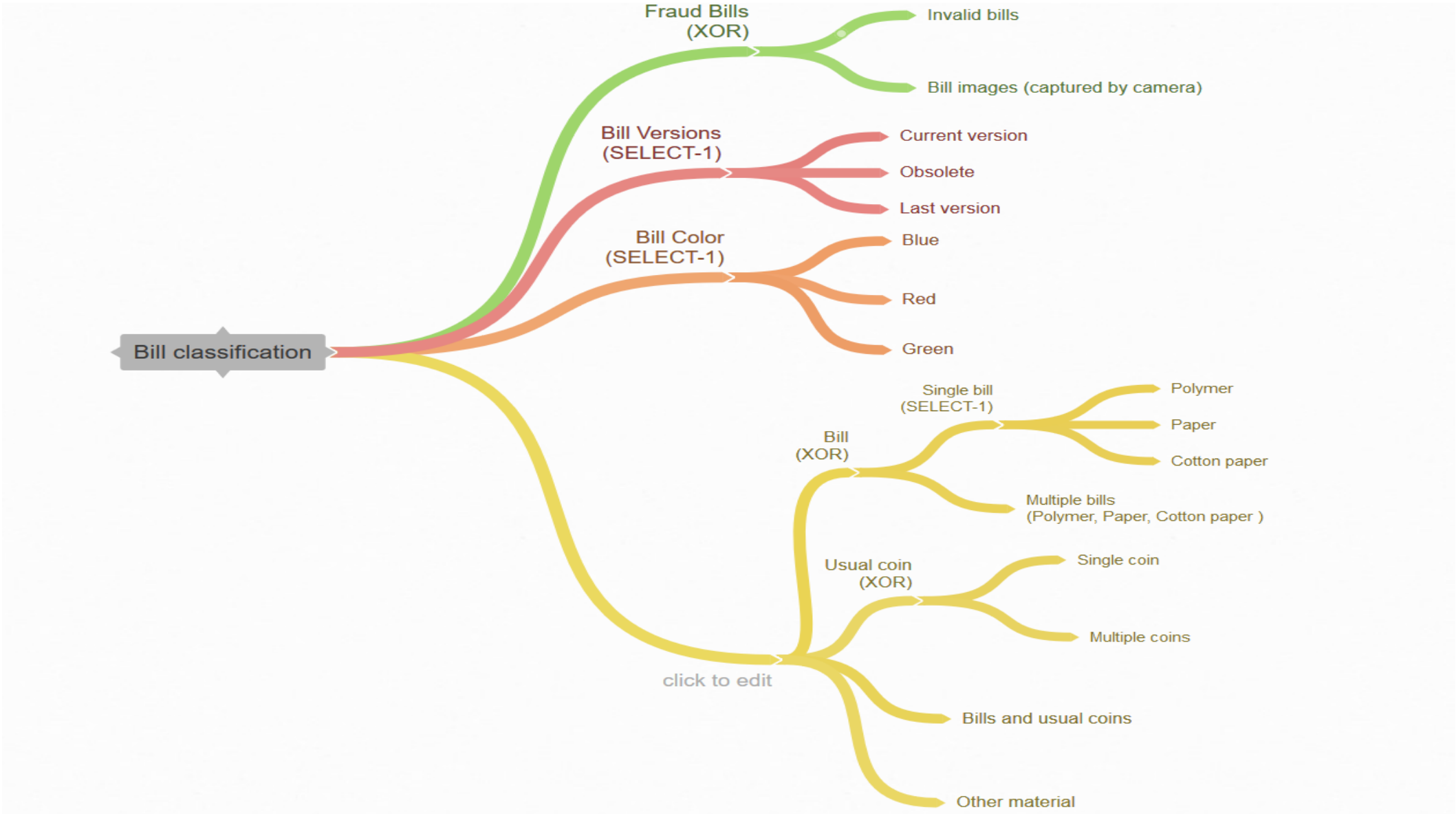
3D Money Reader – Input Tree (Bill State Classification)



Money Reader - AI Function Events



3D Money Reader – Input Tree (Bill Classification)



3D Money Reader – Output Tree

Money Reader - AI Context Decision Table

AI Context Decision Table					Rules																			
Conditions	Equipment type	Ipad	Ipad		F	F	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
			Mini		F	F	–	–	–	–	–	–	–	–	–	–	–	–	T	F	F	F	F	F	F
			Pro		F	F	–	–	–	–	–	–	–	–	–	–	–	–	–	T	F	F	F	F	F
			Air		F	F	–	–	–	–	–	–	I	–	–	–	–	–	–	–	I	F	F	F	F
		Iphone	XS		F	F	–	T	T	T	T	T	–	T	T	T	T	T	F	F	F	F	F	F	F
			S		F	F	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	T	F	F	F
			S		F	F	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	T	T	T
		Mac Laptop			F	T	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
		Dell Laptop			T	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Equipment Position to the bill	Parallel			T	T	F	T	T	T	–	T	F	T	T	T	–	–	–	–	T	T	–	T	T
		Right angle			–	–	F	F	–	–	T	–	T	–	–	–	T	–	–	T	–	–	–	–	–
		Acute			–	–	T	–	–	–	–	–	–	–	–	–	–	T	–	–	–	–	T	–	–
		No gap between bill and camera			–	–	–	–	–	–	–	–	–	–	–	–	–	–	T	–	–	–	–	–	–
	Currencies	Nagiri language	Supported (Indian rupee)		–	–	F	F	F	F	F	T	F	T	F	F	F	F	–	–	–	–	–	–	–
			Unsupported (Nepal rupee)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	T	–	–	–	–	–	–
		Middle Semit language	Supported (UAE Dirham)		–	–	T	T	F	F	F	–	T	–	T	T	F	F	–	–	–	–	–	–	–
			Unsupported (Qatar Riyal)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	T	–	–	–	–	–
		Kanji language	Supported (Japan YEN)		–	–	–	–	T	F	T	–	–	–	–	–	–	–	–	–	–	–	–	–	–
			Unsupported (China Yuan)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	T	–	–	–	–
		Latin - US	Supported (UK Pound)		–	–	–	–	–	T	–	–	–	–	–	–	T	T	–	–	–	T	–	T	–
			Supported (Australia Dollar)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	T
			Unsupported (Vietnam Dong)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	T	–	–
	State	Stable			–	–	T	T	F	F	T	F	F	T	T	T	F	F	T	T	T	F	T	T	T
		Moving			–	–	–	–	T	T	F	T	T	–	–	–	T	T	–	F	–	T	–	F	F
	Distance	Too close (<3 cm)			–	–	F	–	–	–	–	–	–	–	–	–	T	–	T	–	–	–	–	–	–
		Too far (>48 cm)			–	–	F	–	T	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
		Optimal (>= 3 cm & <= 48 cm)			–	–	T	T	–	T	T	T	T	T	T	T	–	–	–	T	T	T	T	T	T
	Environmental Factors	Light Condition	Color light	Blue	–	–	F	–	T	T	F	–	–	–	–	–	–	–	–	–	–	T	–	–	–
				Red	–	–	F	T	–	–	F	–	–	–	–	–	–	–	–	–	–	–	T	–	–
				Green	–	–	F	–	–	–	T	–	–	–	–	–	–	–	–	–	T	T	–	–	–
			White light	Too dark (no light)	–	–	F	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
				Bright light	–	–	F	–	–	–	–	T	–	–	–	–	–	–	–	–	–	–	–	–	–
				Normal light	–	–	T	–	–	–	–	–	T	T	T	–	–	–	T	–	–	–	–	T	T
				Low light	–	–	–	–	–	–	–	–	–	–	–	T	T	T	–	–	–	–	–	–	–
		Background	No Noisy objects		–	–	T	T	T	T	F	F	F	F	T	T	F	F	T	T	T	–	T	T	T
			Noisy objects	Single object	–	–	–	–	–	–	T	T	T	F	–	–	T	T	–	–	–	–	–	–	–
				Multiple objects	–	–	–	–	–	–	–	–	–	T	–	–	–	–	–	–	–	T	–	–	–
			Different surfaces	Curved	–	–	F	F	F	F	T	F	F	F	F	F	T	F	–	–	–	–	–	–	–
				Rough	–	–	–	–	T	T		T	–	–	–	T	–	–	T	T	T	–	–	–	–
				Plain	–	–	T	T	–	–	–	–	T	T	T	–	–	T	–	–	–	T	T	T	T
Actions	Money Reader cannot be installed on the machine				T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
	Money Reader can read the value of the bill correctly in conrect selected language and currency type.				–	T	T	T	F	T	F	F	T	–	T	T	F	F	F	F	F	T	F	T	T
	Money Reader cannot recognize the bill and cannot read the value of the bill				–	–	–	–	T	–	T	T	–	–	–	–	T	T	T	T	T	–	T	–	–
	Money Reader can read the value of only the bill correctly in correct selected language and currency type, cannot recognize and cannot read the value of the coins.				–	–	–	–	–	–	–	–	–	T	–	–	–	–	–	–	–	–	–	–	–

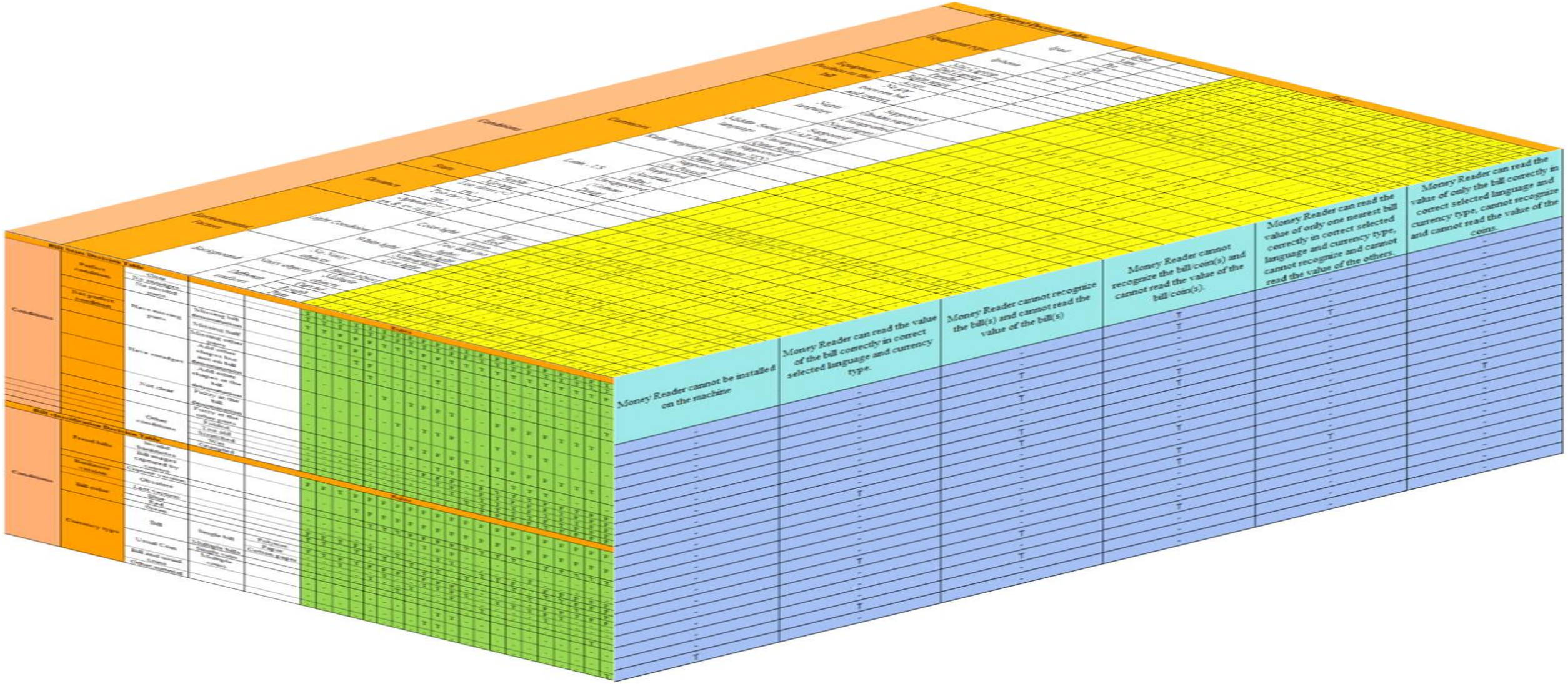
Money Reader - Bill State Decision Table (State Classification)

Bill State Decision Table					Rules																				
Conditions	Perfect condition	Clear			T	T	T	T	T	T	T	F	F	F	F	T	F	F	F	F	F	F	F	F	T
		No smudges			T	T	T	T	T	F	F	T	T	T	F	F	T	T	T	F	T	T	T	T	T
		No missing parts			T	T	F	F	F	T	T	T	F	F	T	T	T	T	T	T	T	T	T	T	F
	Not perfect condition	Have missing parts	Missing bill denomination		-	-	T	F	F	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-
			Missing half		-	-	-	T	F	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-
			Missing other parts		-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T
		Have smudges	Add other shapes but not on bill denomination		-	-	-	-	-	T	-	T	F	F	T	-	-	F	F	F	F	T	T	T	-
			Add other shapes at the bill denomination		-	-	-	-	-	-	T	-	T	T	F	-	-	T	T	T	T	-	-	-	-
		Not clear	Fuzzy at the bill denomination		-	-	-	-	-	-	-	T	T	F	F	T	-	T	F	F	F	T	T	T	-
			Fuzzy at the other parts		-	-	-	-	-	-	-	-	-	T	T	-	-	-	T	T	T	-	-	-	-
		Other conditions	Folded		-	-	-	-	-	-	-	-	F	F	F	-	F	T	T	T	F	F	T	F	F
			Too old		-	-	-	-	-	-	-	-	-	T	F	-	F	T	F	F	F	T	F	F	F
			Scratched		-	-	-	-	-	-	-	-	-	-	T	-	T	T	T	F	T	F	T	F	F
			Wet		-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	F	T	F	F	F
			Crumpled		-	-	-	-	-	-	-	-	-	-	-	-	T	F	F	F	T	F	F	F	F
Actions	Money Reader can read the value of the bill correctly in correct selected language and currency type.				T	T	T	T	T	T	T	T	T	T	T	T	F	T	T	T	F	T	T	T	T
	Money Reader cannot recognize the bill and cannot read the value of the bill				-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	T	-	-	-	-

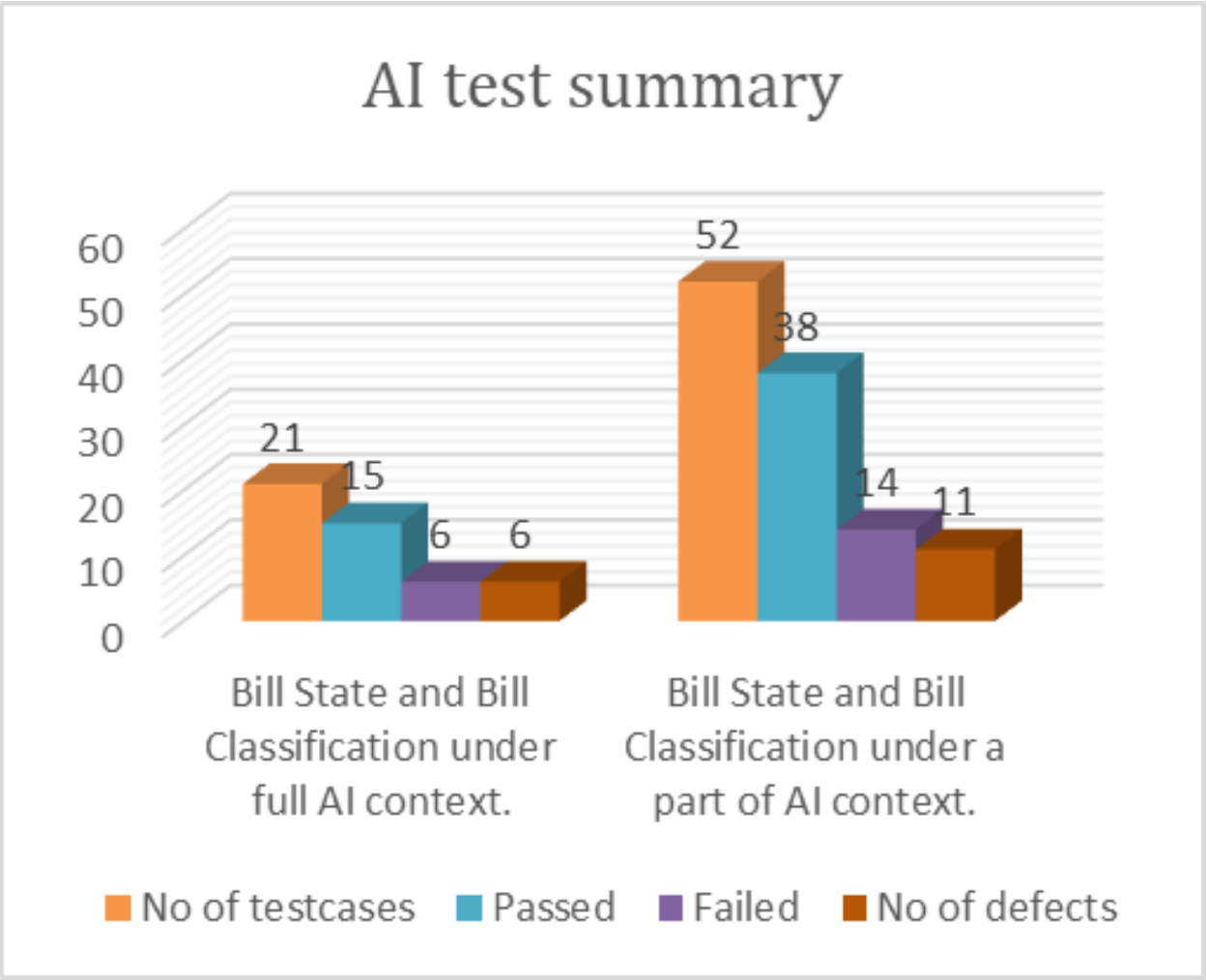
Money Reader - Bill State Decision Table (Bill Classification)

Bill classification Decision Table					Rules																				
Conditions	Fraud bills	Invalid banknotes			F	F	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
		Bill images captured by camera			-	-	-	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
	Banknote version	Current version			-	-	-	-	T	T	T	F	F	F	F	T	T	T	T	-	T	T	T	T	T
		Obsolete			-	-	-	-	-	-	-	T	T	F	F	-	-	-	-	-	-	-	-	-	-
		Last version			-	-	-	-	-	-	-	-	-	T	T	-	-	-	-	T	-	-	-	-	-
	Bill color	Blue			F	F	T	F	T	-	T	-	-	-	-	-	T	T	F	-	F	F	F	F	F
		Red			T	T	-	T	-	T	-	T	-	T	-	T	-	-	T	-	T	T	F	F	F
		Green			-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	T	-	-	T	-	F
	Currency type	Bill	Single bill	Polymer	T	T	F	T	F	F	F	-	F	F	F	-	-	-	-	-	F	F	T	F	F
				Paper	-	-	T	-	-	T	-	-	T	T	F	T	-	T	T	T	F	T	-	-	-
				Cotton paper	-	-	-	-	T	-	-	T	-	-	F	-	-	-	-	-	T	-	-	-	-
			Multiple bills		-	-	-	-	-	-	T	-	-	-	-	-	T	-	-	-	-	-	-	-	-
		Usual Coin	Single coin		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-
			Multiple coins		-	-	-	-	-	-	-	-	-	T	T	-	-	-	-	-	-	-	-	-	-
		Bill and usual coins			-	-	-	-	-	-	-	-	T	T	-	-	-	-	-	-	-	-	-	-	-
		Other material			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T
Actions	Money Reader can read the value of the bill correctly in correct selected language and currency type.				T	T	F	F	T	T	F	F	F	F	F	T	F	T	T	T	T	T	T	F	F
	Money Reader cannot recognize the bill/coin(s) and cannot read the value of the bill/coin(s).				-	-	T	T	-	-	F	T	T	F	T	-	F	-	-	-	-	-	-	T	T
	Money Reader can read the value of only one nearest bill correctly in correct selected language and currency type, cannot recognize and cannot read the value of the others.				-	-	-	-	-	-	T	-	-	F	-	-	T	-	-	-	-	-	-	-	-
	Money Reader can read the value of only the bill correctly in correct selected language and currency type, cannot recognize and cannot read the value of the coins.				-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-

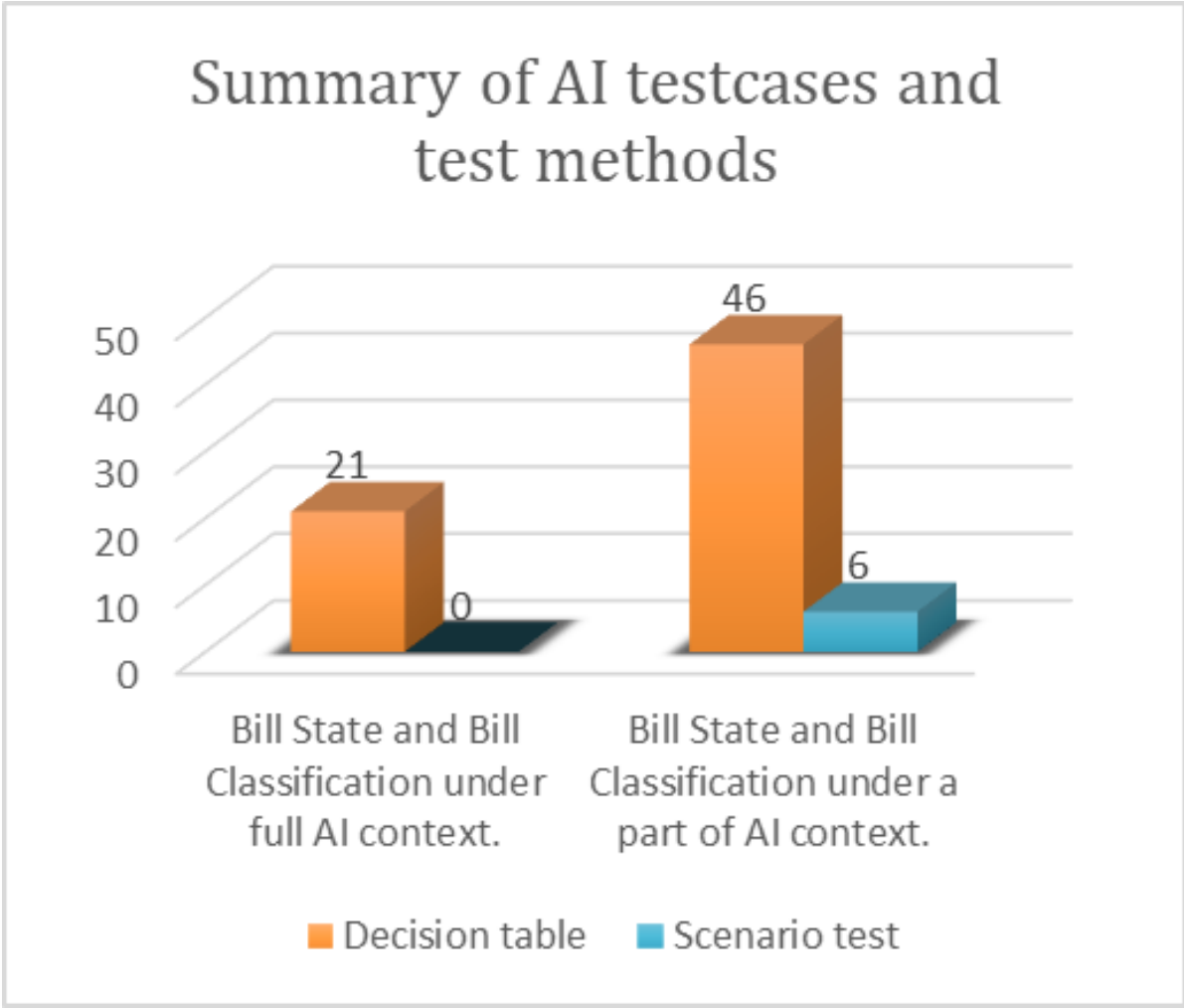
Money Reader - 3D Decision Table



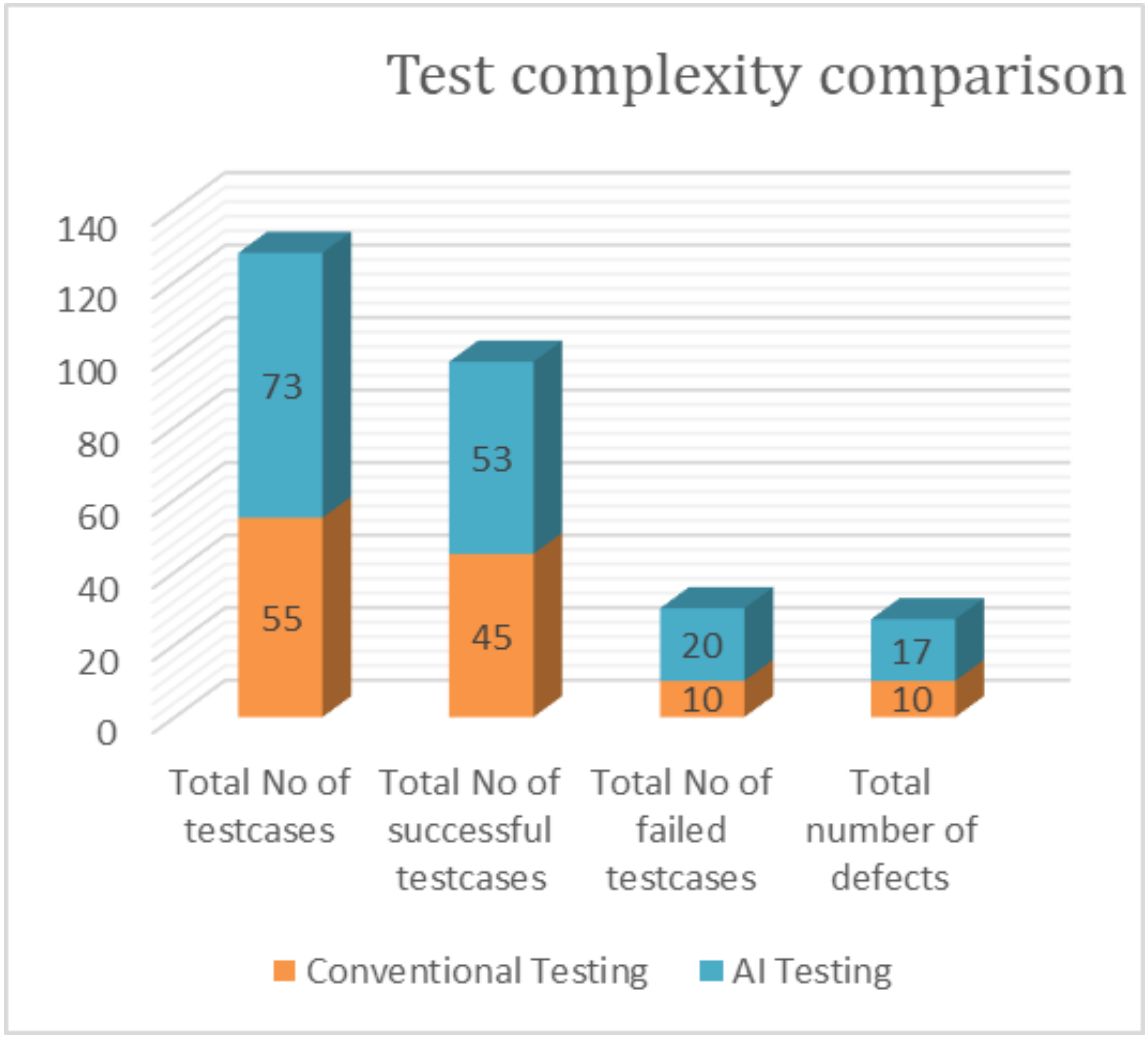
Money Reader - AI Test Summary



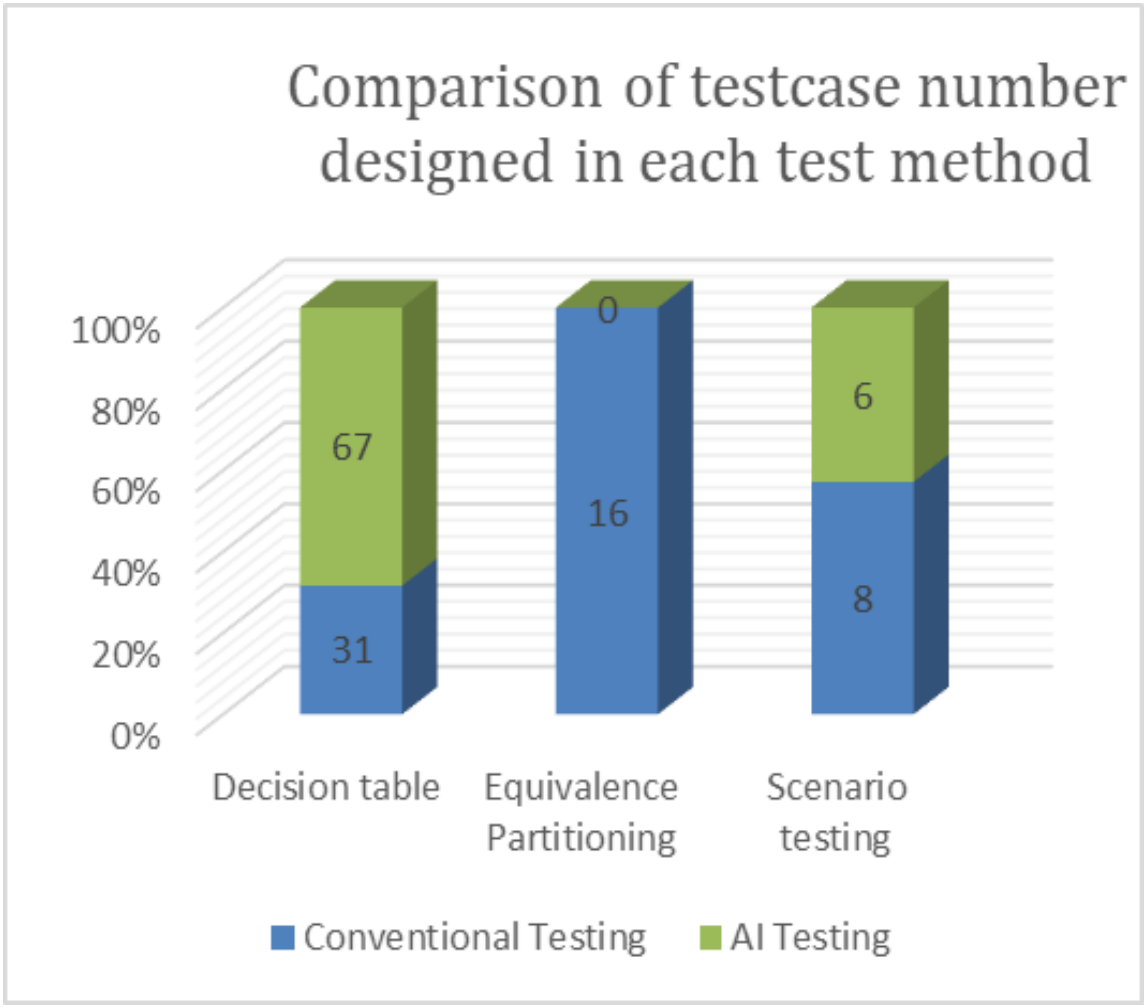
Summary of AI Testcases and Test Methods.

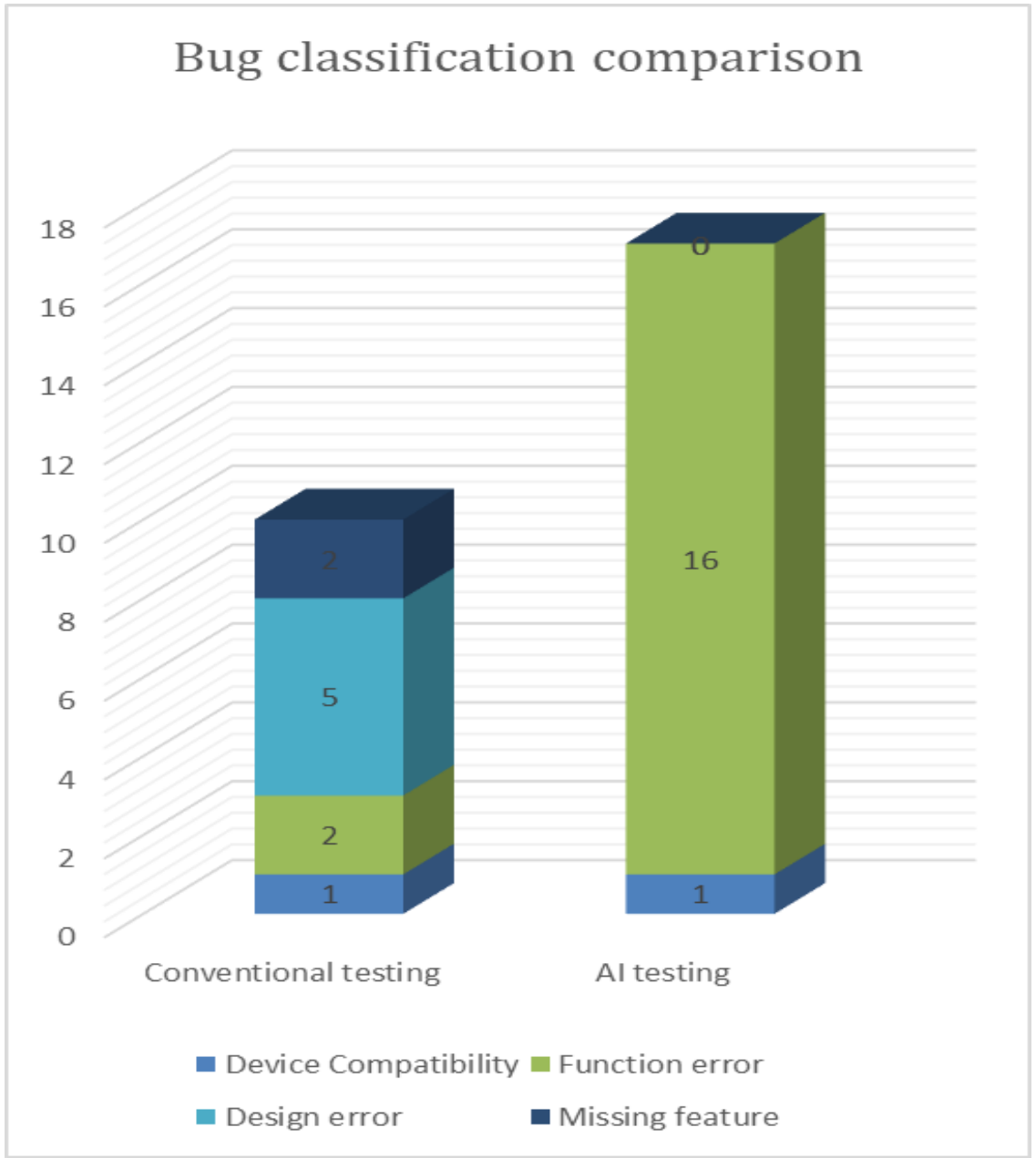


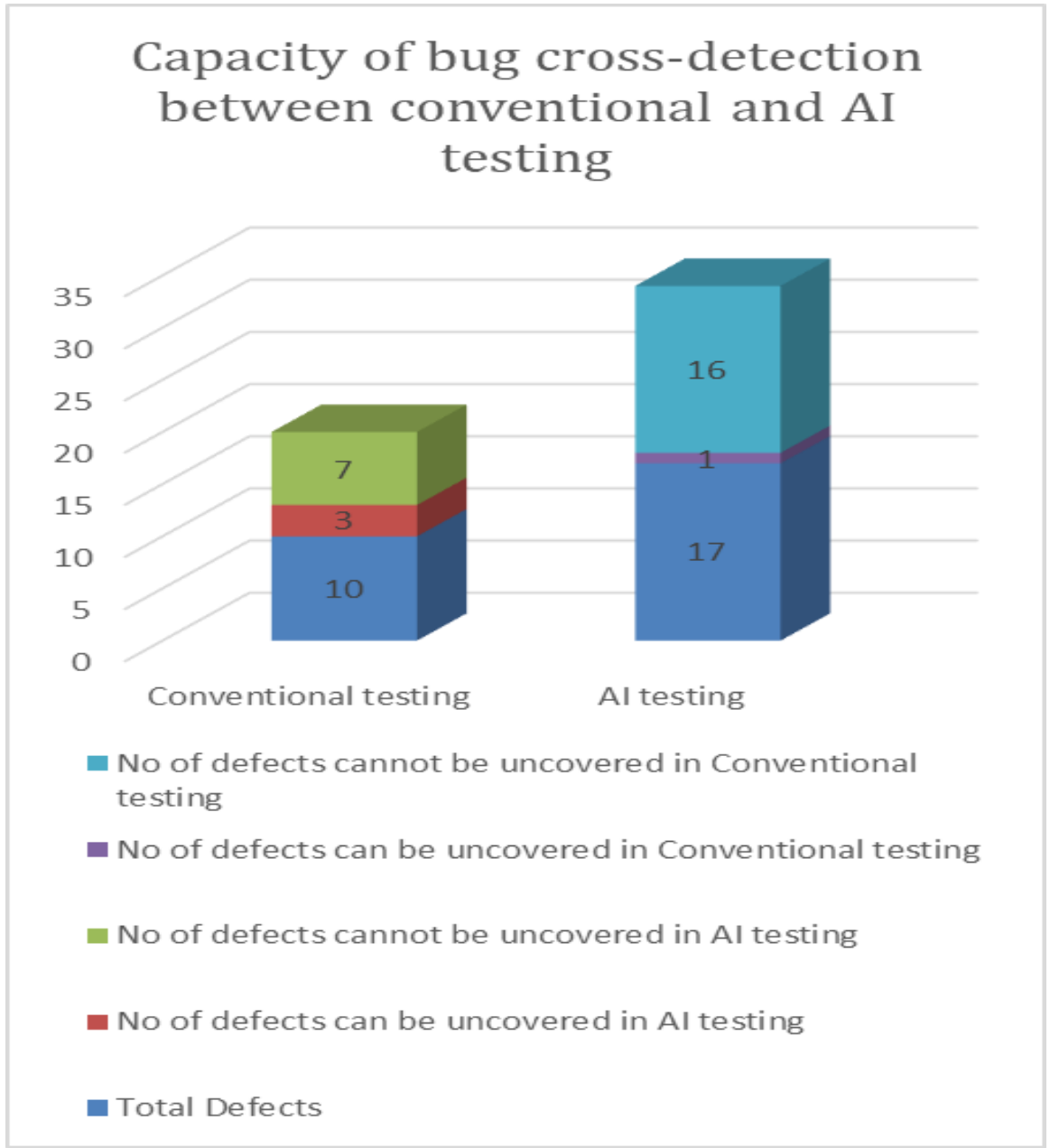
Test Complexity Comparison Between Conventional Testing and AI Testing.



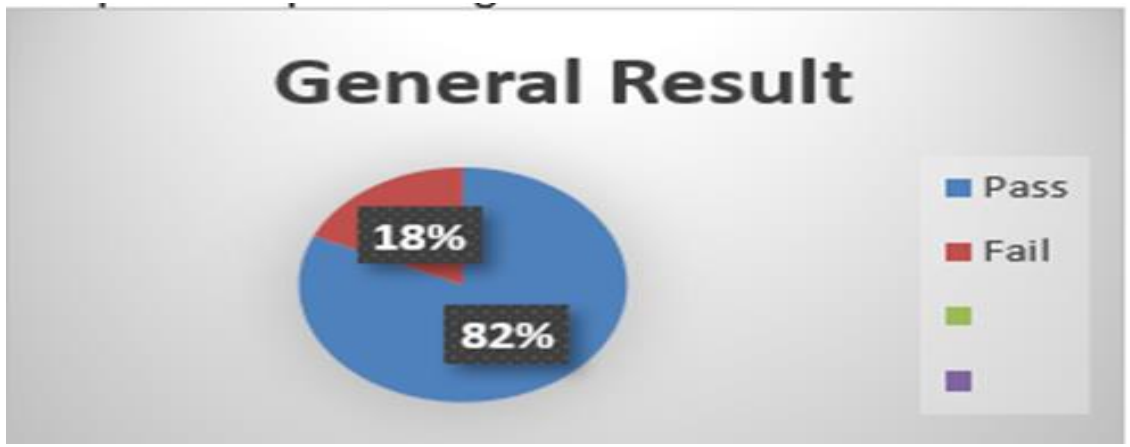
A Comparison of Testcase Number Designed in Each Test Method.

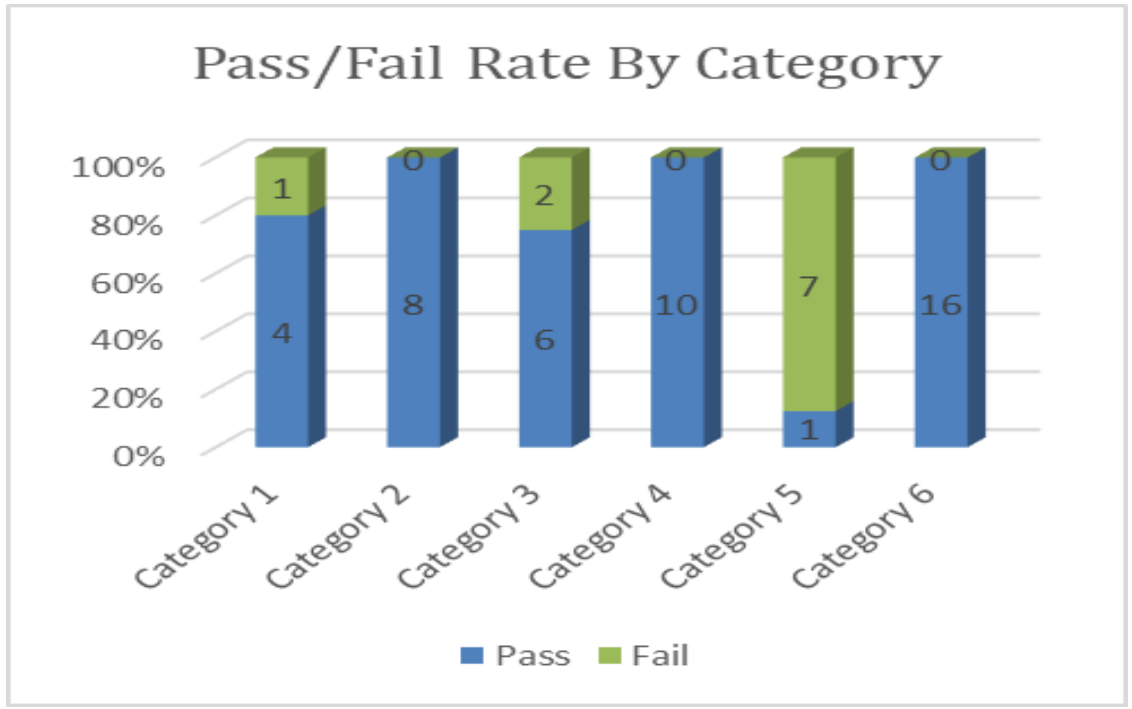



Bug classification comparison
18
16
14
12
10
8
6
4
2
0
2
5
2
1
0
16
1
Conventional testing
AI testing
Device Compatibility
Function error
Design error
Missing feature

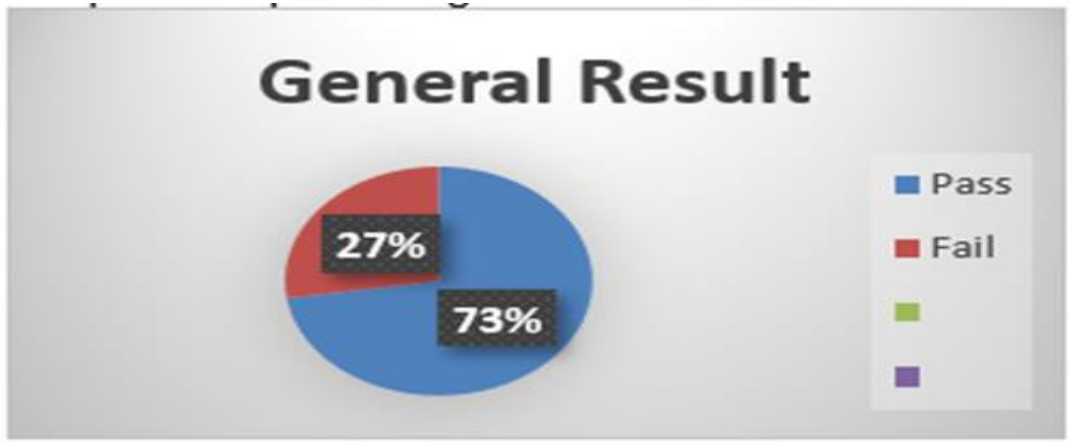

Capacity of bug cross-detection between conventional and AI testing
35
30
25
20
15
10
5
0
7
3
10
16
1
17
Conventional testing
AI testing
No of defects cannot be uncovered in Conventional testing
No of defects can be uncovered in Conventional testing
No of defects cannot be uncovered in AI testing
No of defects can be uncovered in AI testing
Total Defects

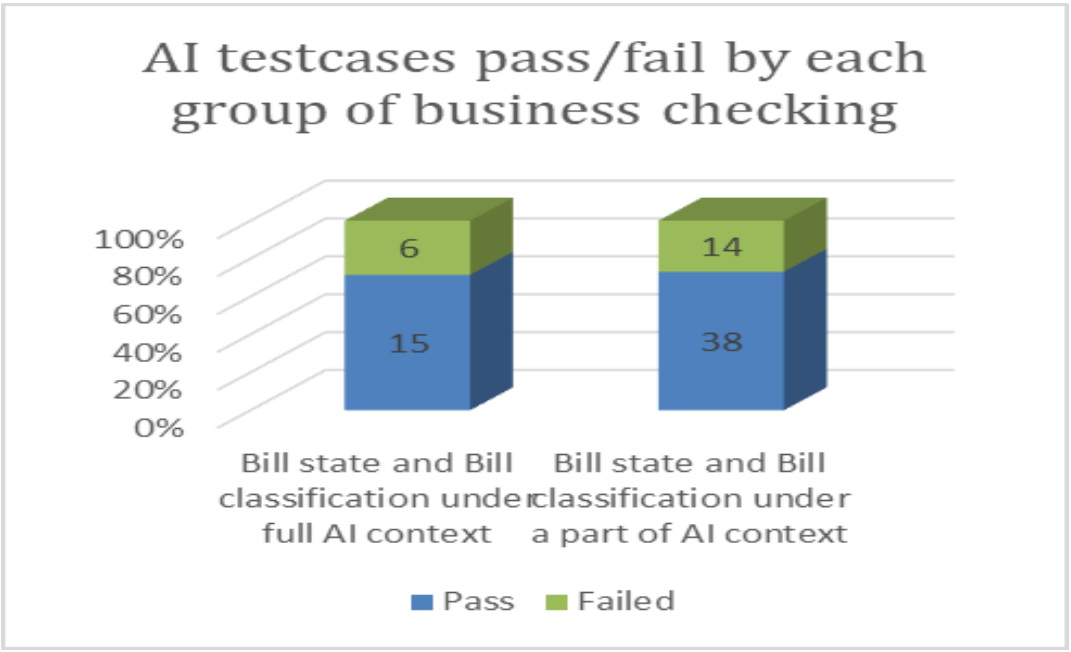
Conventional General Test Result



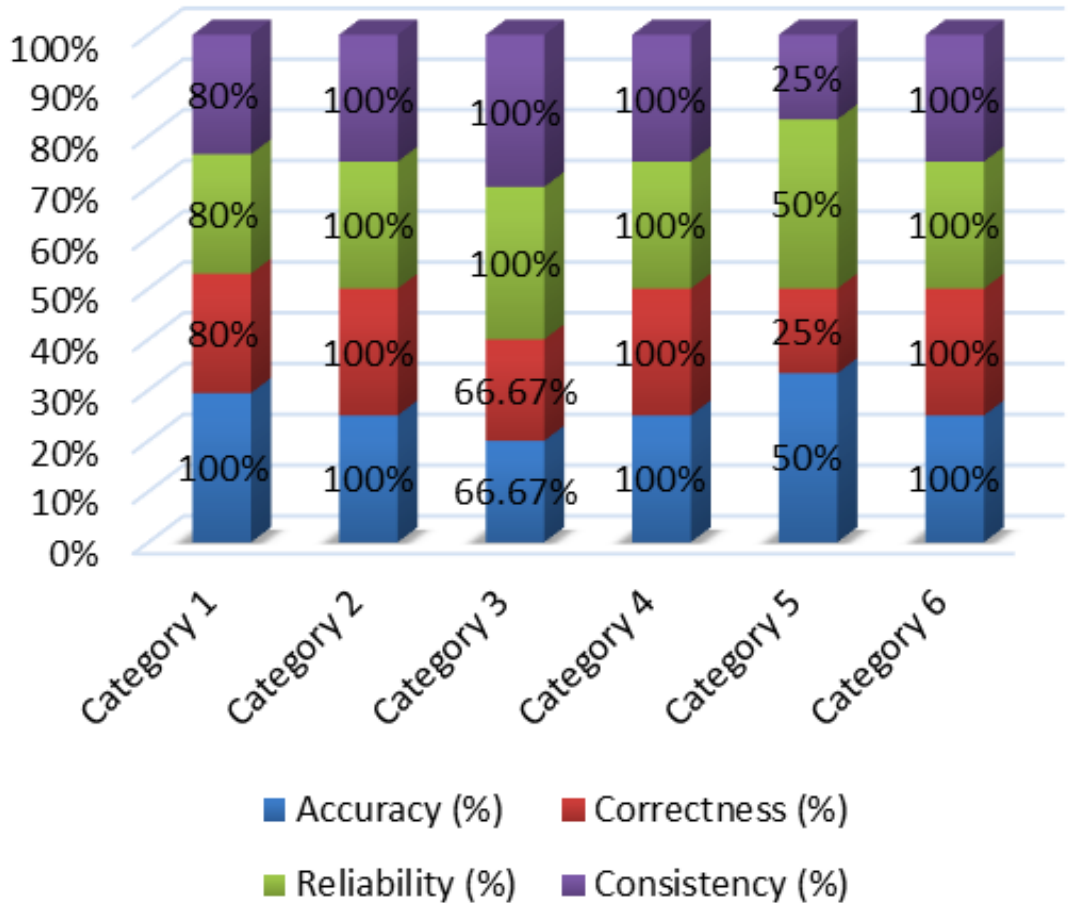


AI General Test Result

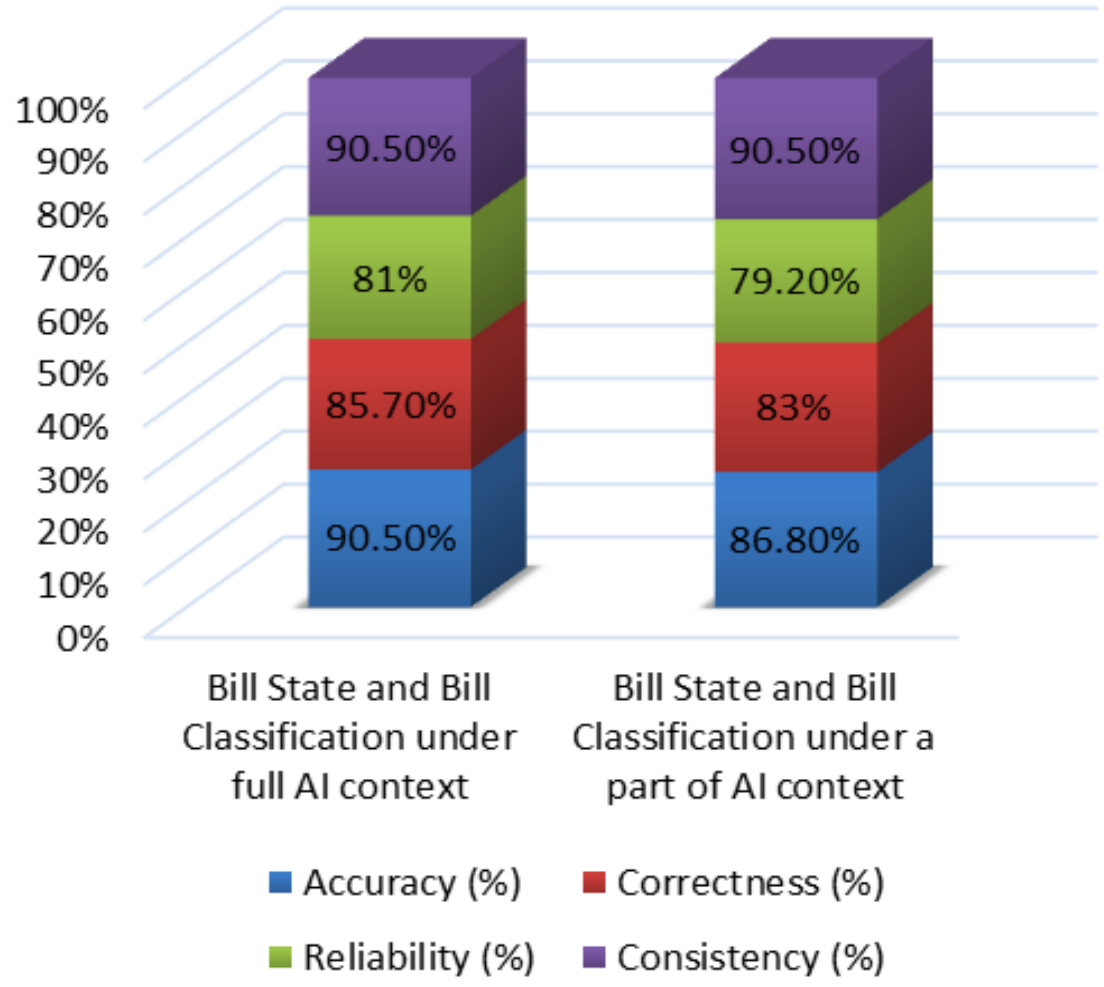




Conventional quality assurance results with metrics and assessments



AI quality assurance results with metrics and assessments



AI System Testing – A Case Study – Google Allo

About Google Allo

- Messaging Application
- Embedded Google Assistant
- Ability to chat in text, emojis, images, GIFs, Memes, etc
- Use Google Assistant in messages
- Smart Reply Suggestion
  - Replies also based on personality
- Image recognition

AI System Testing – A Case Study – Google Allo

The following are the areas of the application that we cover:
1. Start Chat feature of Google Allo
2. Send Message feature of Google Allo
3. QR code feature of Google Allo
4. Chat Setting feature of Google Allo

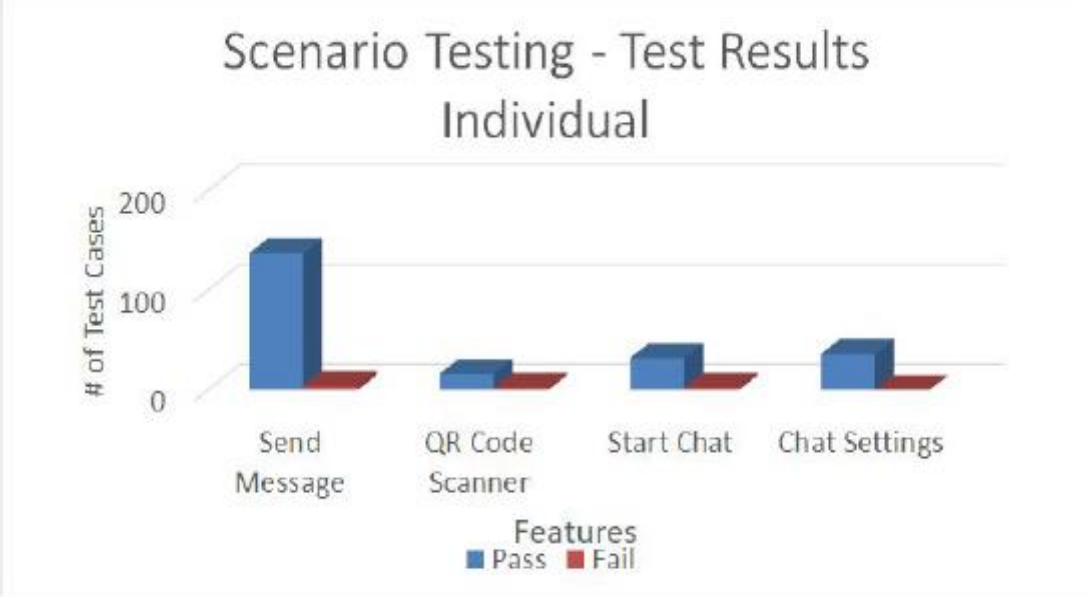


Fig 2.1.2.2a. Scenario Testing Results

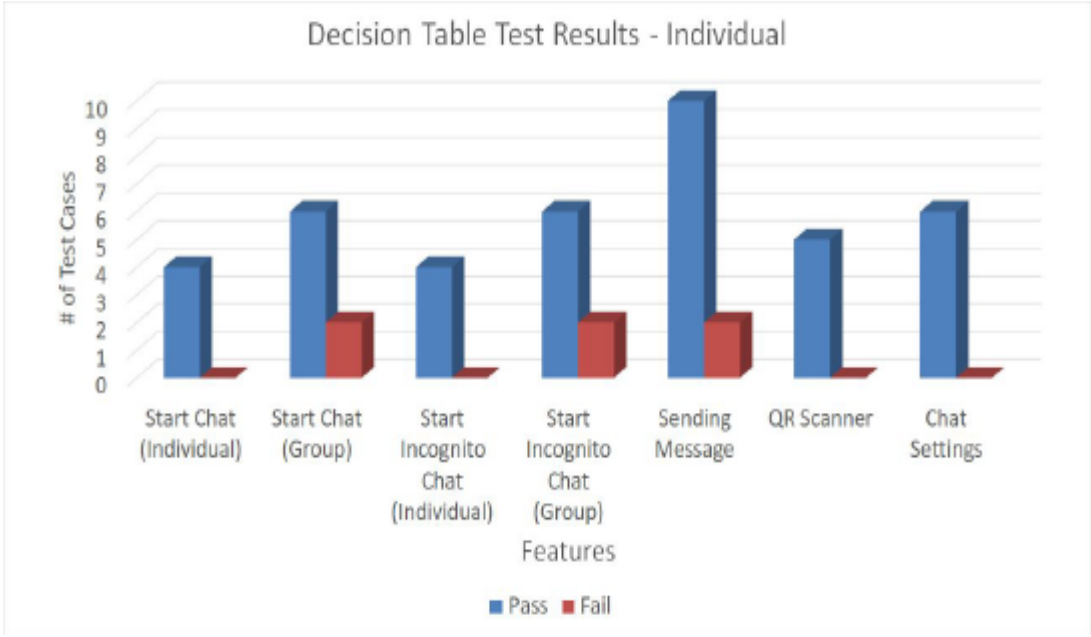
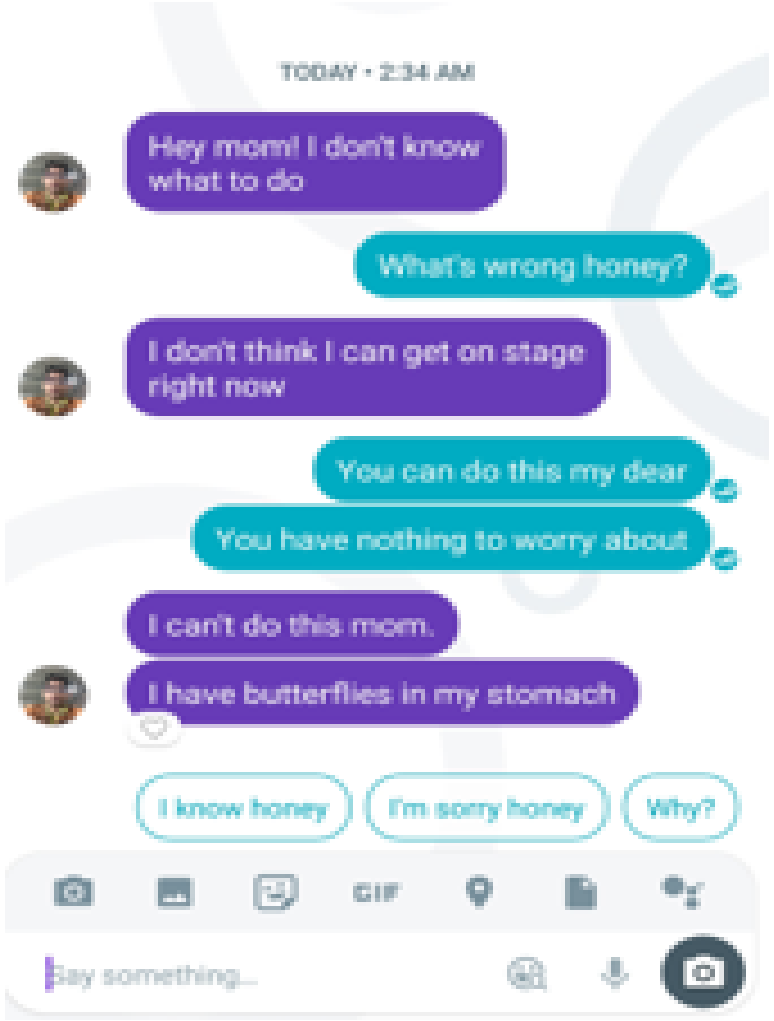


Fig 2.1.1.2A: Decision Table Test Results

AI System Testing – A Case Study – Google Allo

AI Features

- Text Recognition
- Emoji Recognition
- Smart Reply Suggestion



AI System Testing – A Case Study – Google Allo

Figure 3.4e. Passed test for multiple smilies

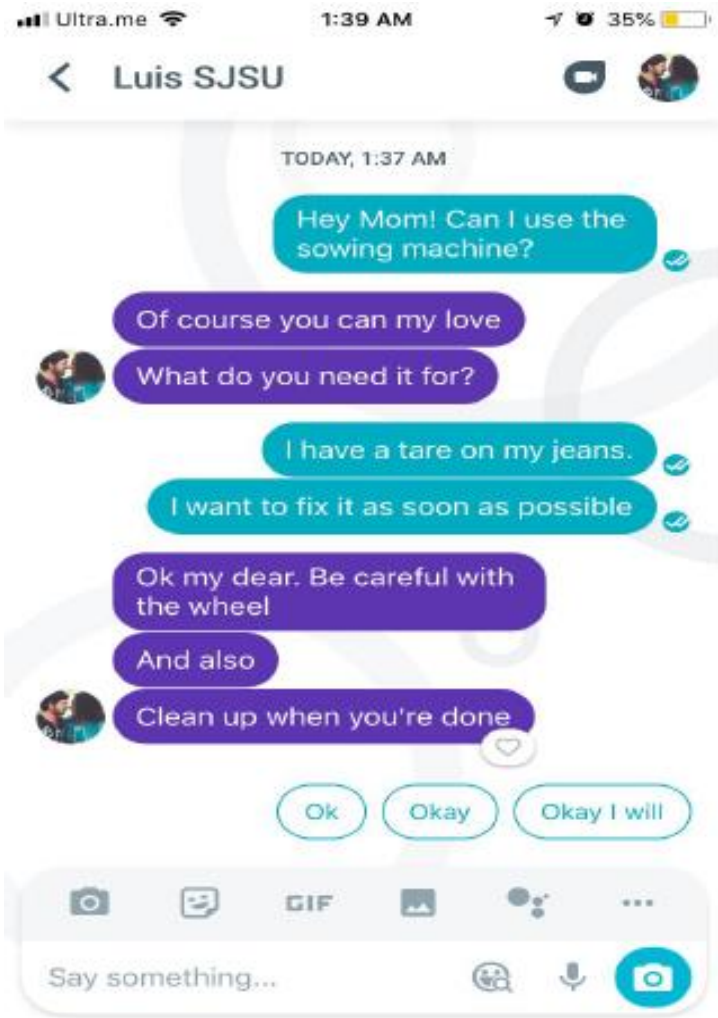


Figure 3.4f. Passed test case for Imperative comlete

Figure 3.4g. Passed test case for Exclamatory sentence

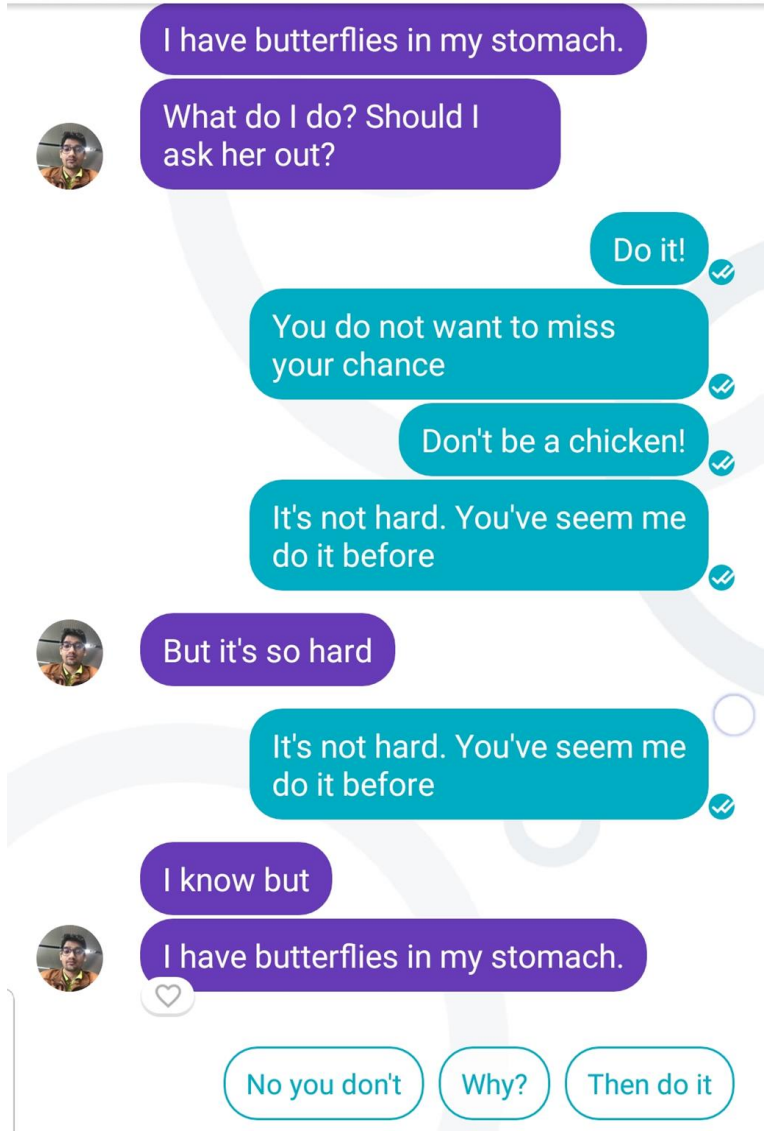
AI System Testing – A Case Study – Google Allo

Figure 3.4i. Failed test case for Spanish language

Explanation:

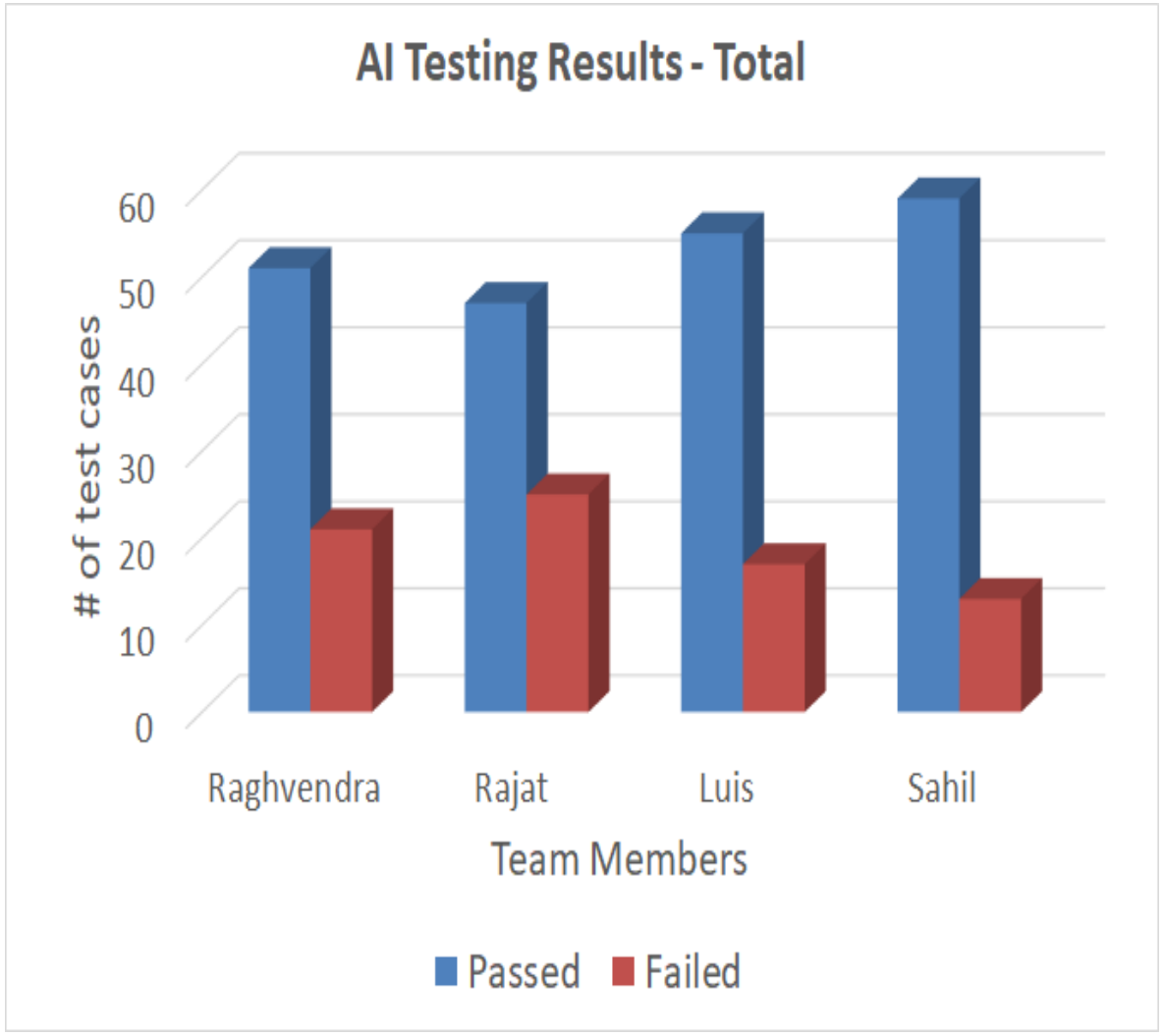
In this chat, we see 1 person asking another how he is doing and other telling him that he punched a guy in face. When first person responded with *Pelea!* , we expected to see something related to fight, but only one suggestion was relevant. We considered it a fail for 2 reason, first being just one relevant suggestion and other one is that it gave 2 similar suggestion, which should not be the case.

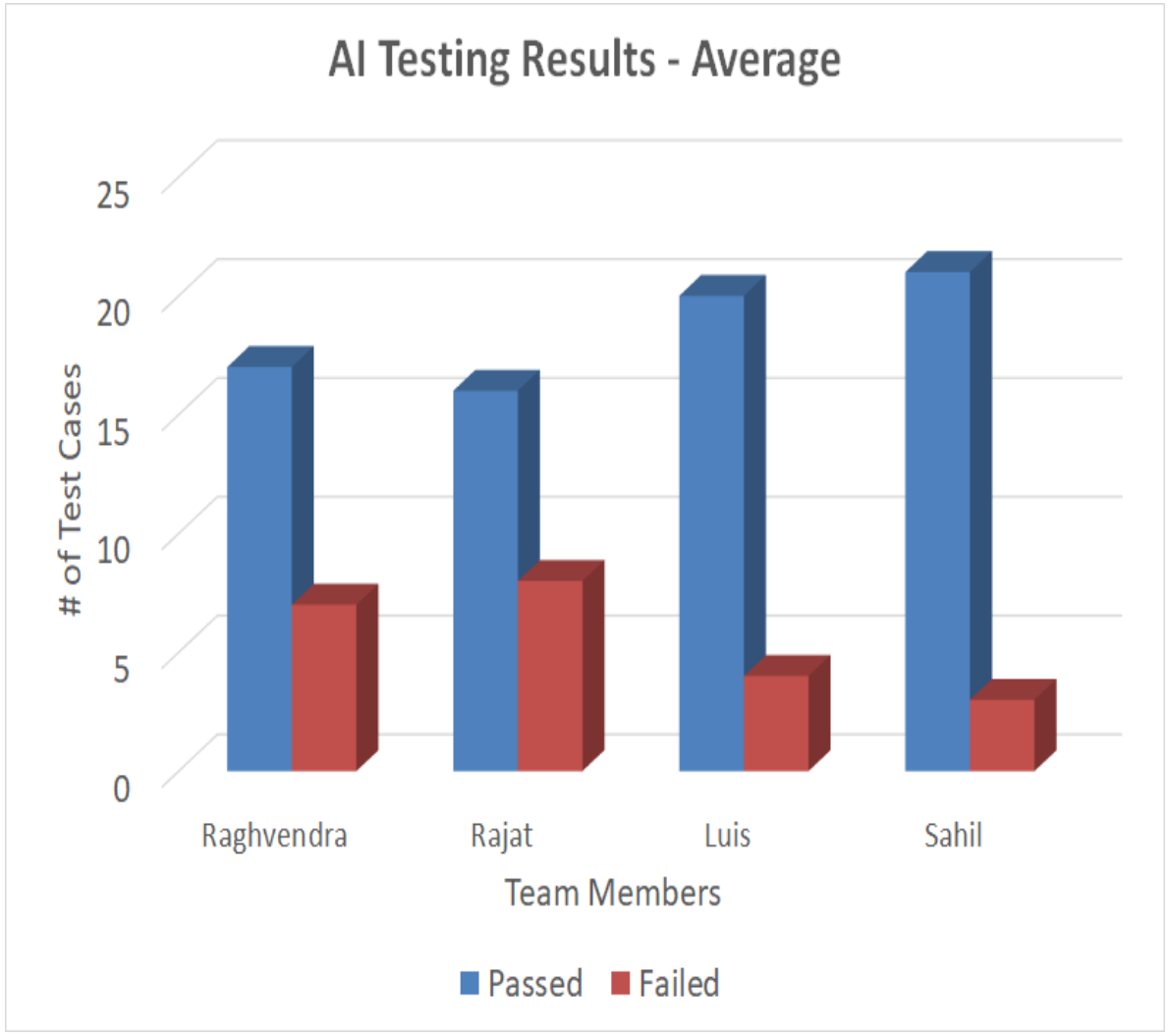
Reason for not enough relevant suggestion could be that AI was not trained enough for Spanish which we could see with so many test getting failed.



AI System Testing – A Case Study – Google Allo

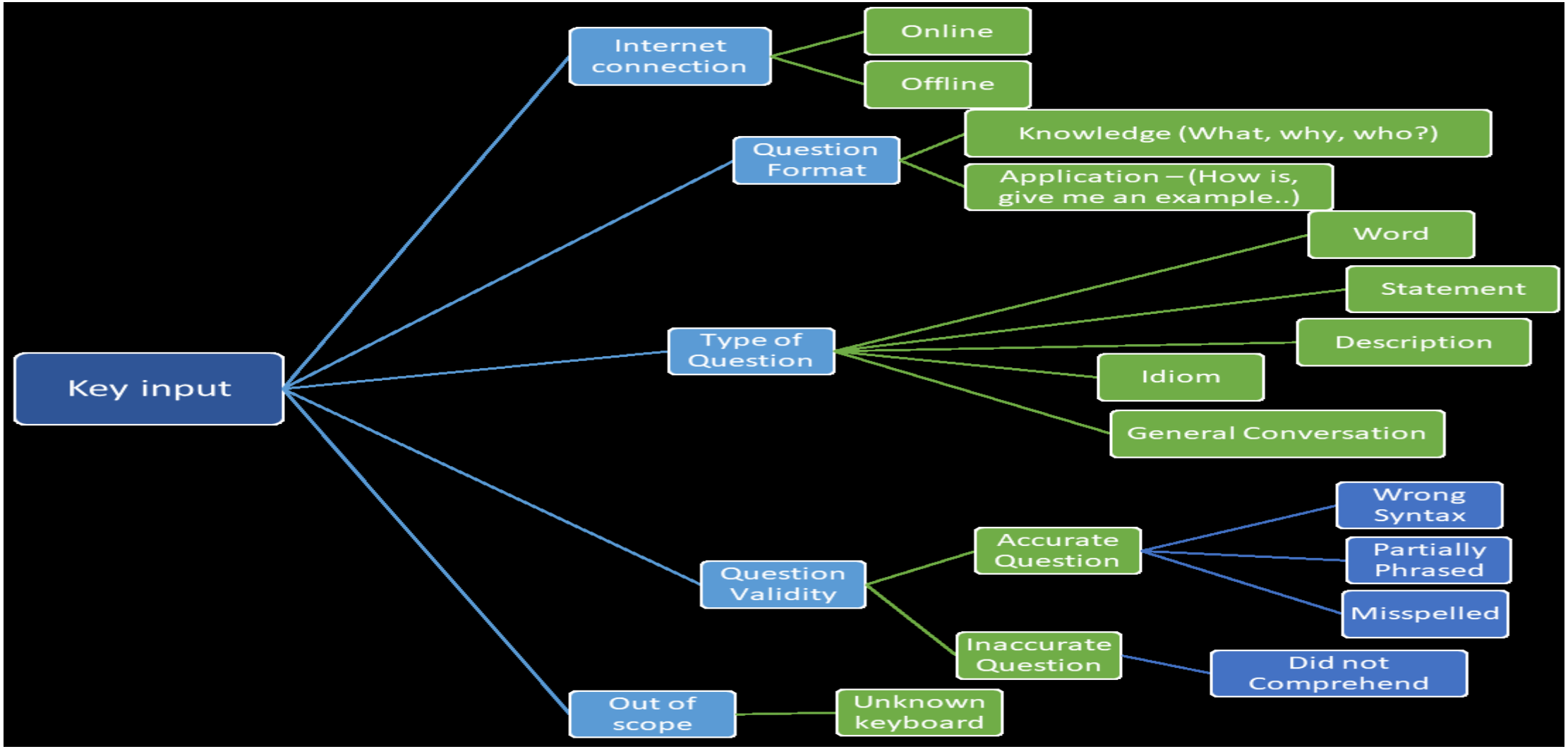
AI Testing - Results





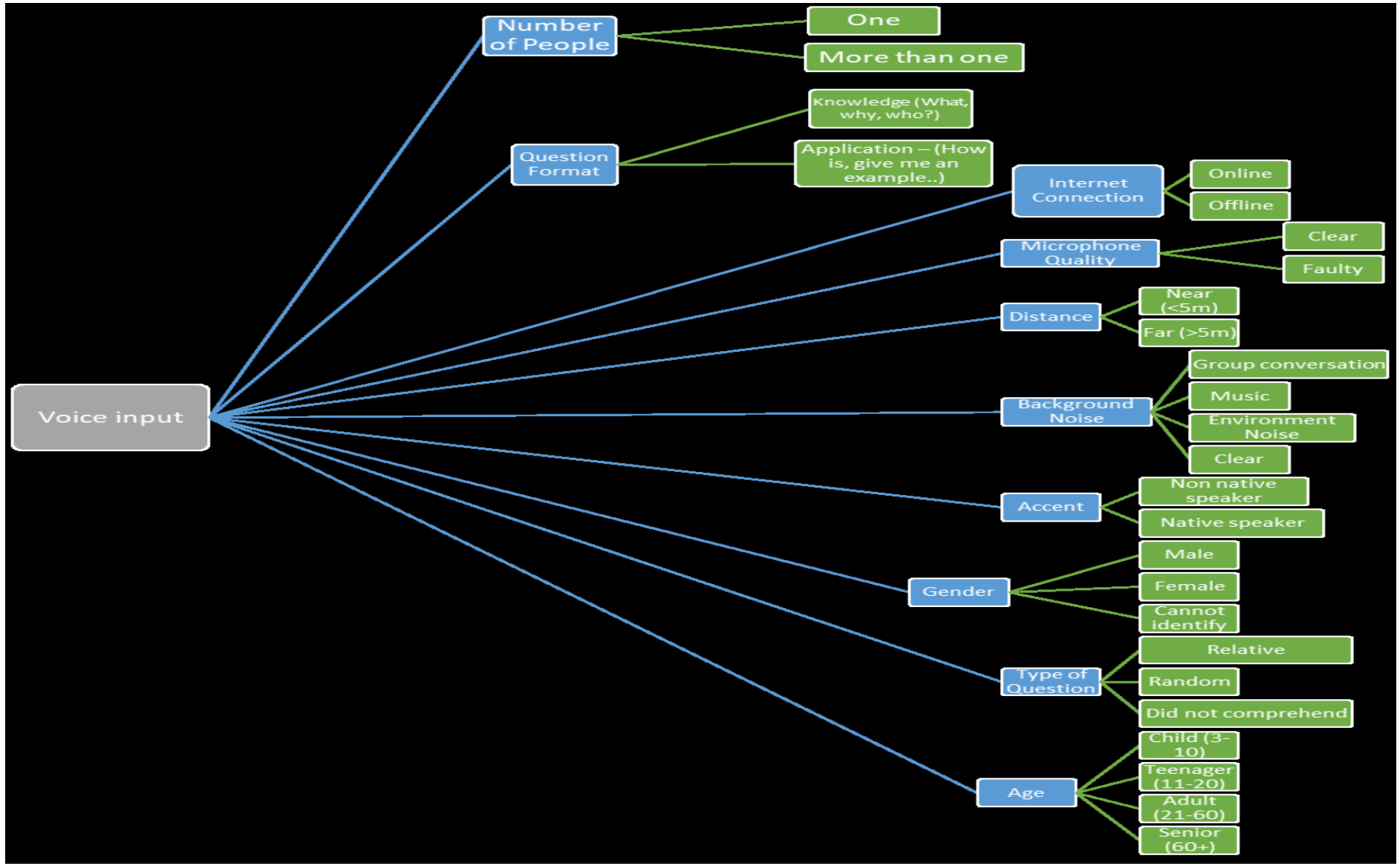
AI System Testing – A Case Study – Google Allo

AI Testing Model



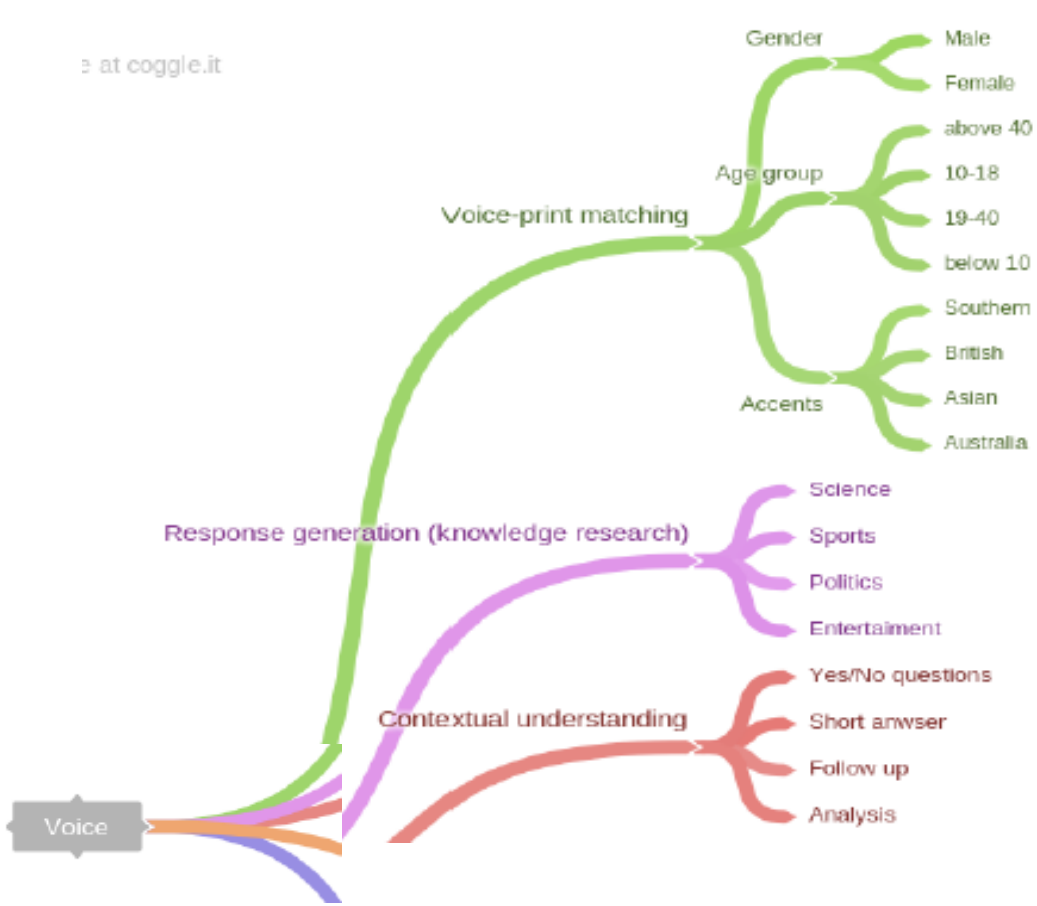
AI System Testing – A Case Study – Google Allo

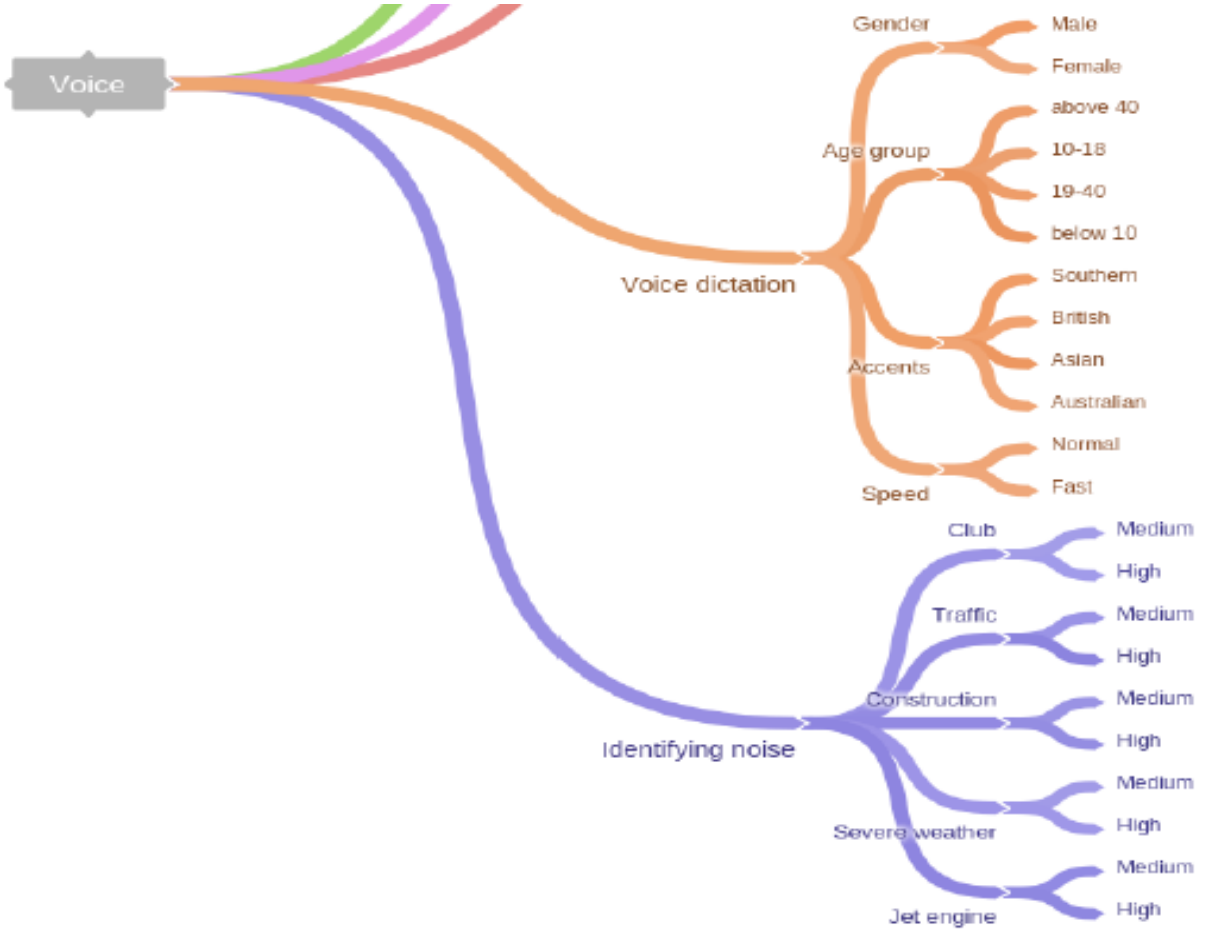
AI Testing Model



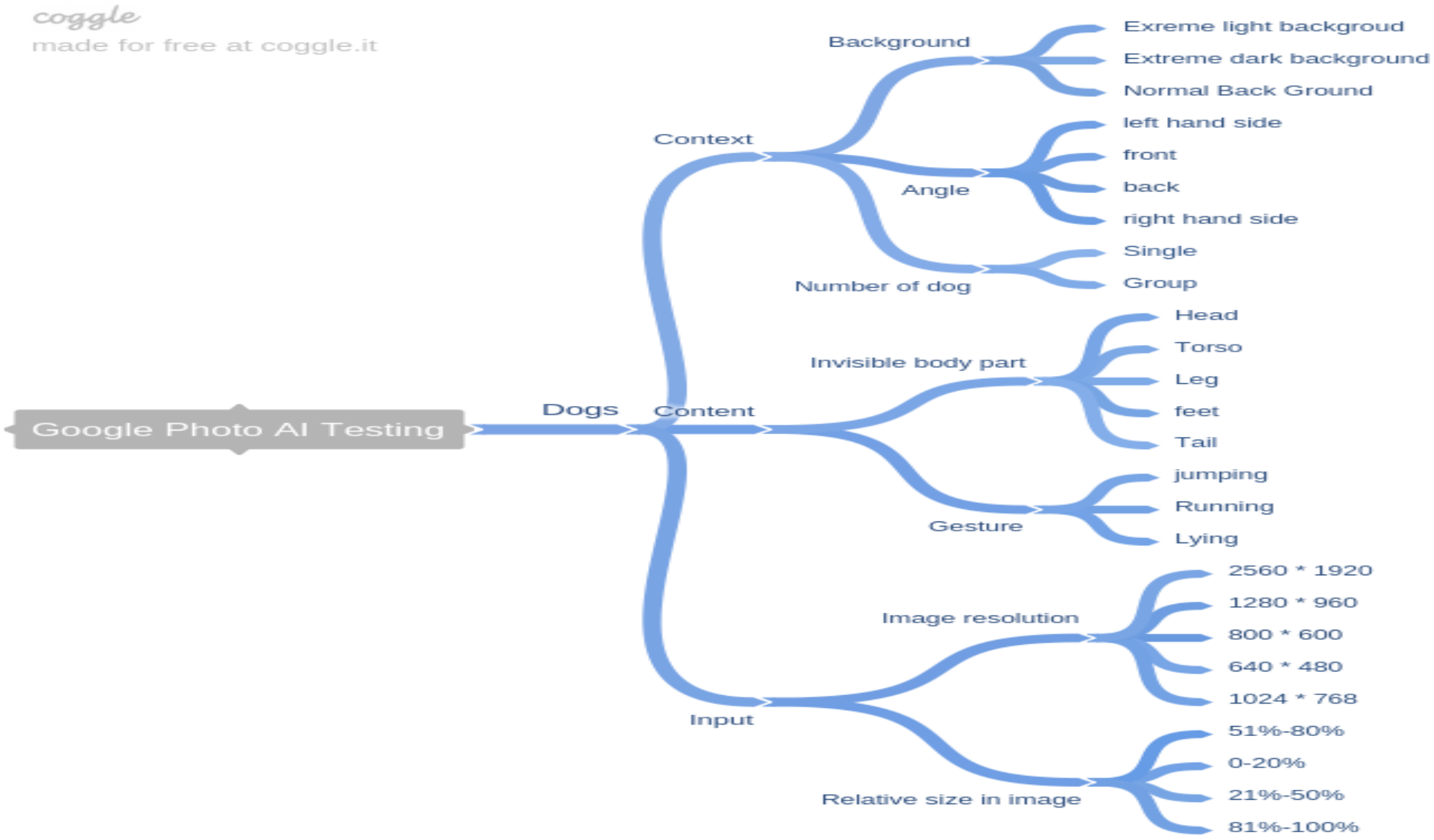
AI System Testing – A Case Study – Google Allo

AI Testing Model





AI System Testing – A Case Study – Google Photo



AI System Testing – A Case Study – Google Allo

Observations:

- There needs to be context for smart replies to become relevant
- Smart reply takes literal meaning of Idioms
- Emojis are better recognized with text attached to them
- English and Hindi had better performance than Spanish
- Onomatopoeias are not recognized

Experience and Lessons Learned:

- AI testing is more challenging than conventional testing
- Difficult to cover all possible contexts, content, external factors
- Automated data generating tool would make AI testing relatively easier to execute
- Proper training and context are the prerequisites for smart replies
- Still a long way to go for AI