Mutation testing for DSLs

The case of task-oriented chatbots

Pablo Gómez-Abajo

Modelling & Software Engineering Research Group

Universidad Autónoma de Madrid

LangDev
CON 2024

Seville 17-19 October, 2024

https://langdevcon.org

Introduction

630I

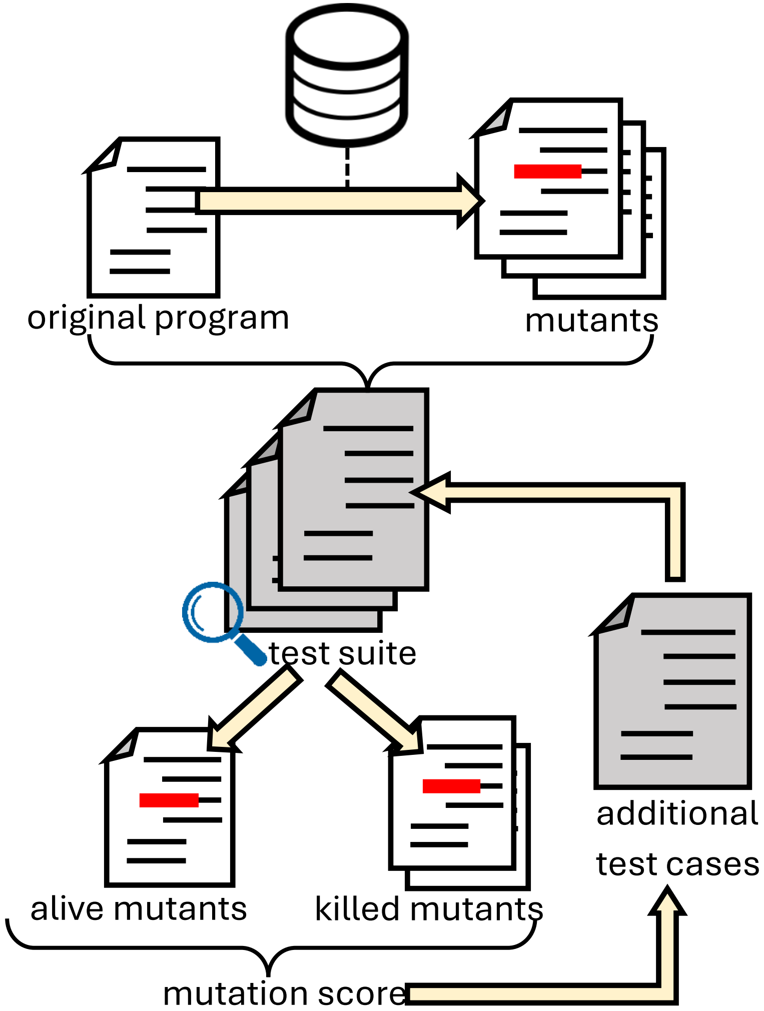
ProxyHands

- DSLs are increasingly used to solve problems in specific domains
- Like any other programming language, DSLs need to be tested
  - Usually by creating and using test suites
- Mutation testing (MuT) is a common technique used to improve such software test suites quality

What is mutation testing?

630I

ProxyHands



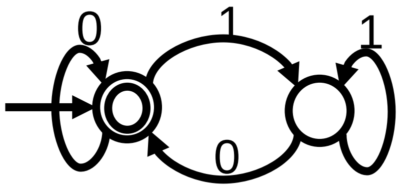
- Approach of software testing to assess the quality of the test suites
- Injection of syntax changes in a program by using a set of mutation operators
- The mutations introduced emulate common programming faults
- Useful to improve the quality of the test suites and the mutation operators set

630I

ProxyHands

Mutation testing for automata

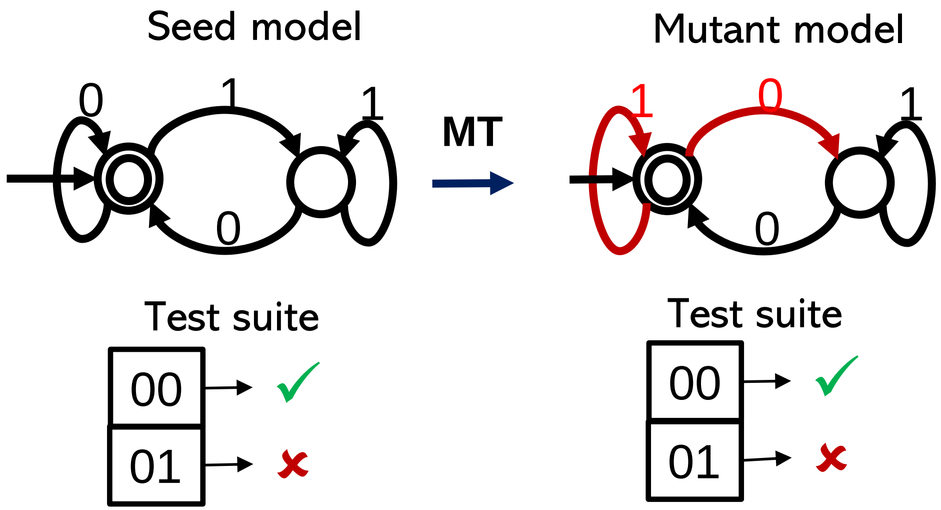
Seed model



Test suite

00	✓
01	✗

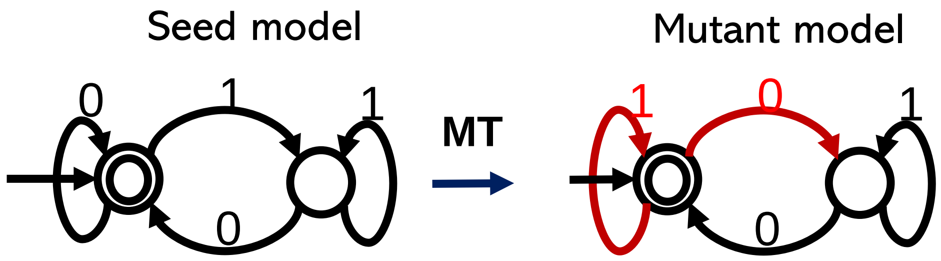
Mutation testing for automata



Mutation testing for automata

630I

ProxyHands



Test suite

00 →	✓
01 →	✗

Test suite

00 →	✓
01 →	✗

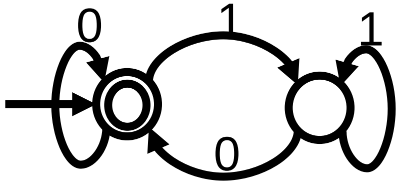
The mutant is alive ✗

630I

ProxyHands

Mutation testing for automata

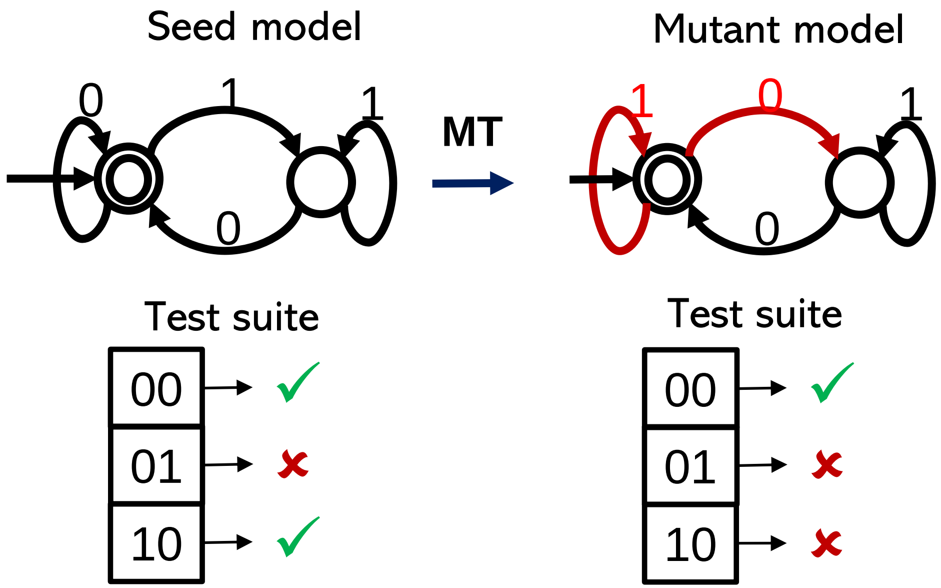
Seed model



Test suite

00	→	✓
01	→	✗
10	→	✓

Mutation testing for automata



Test suite

00	✓
01	✗
10	✓

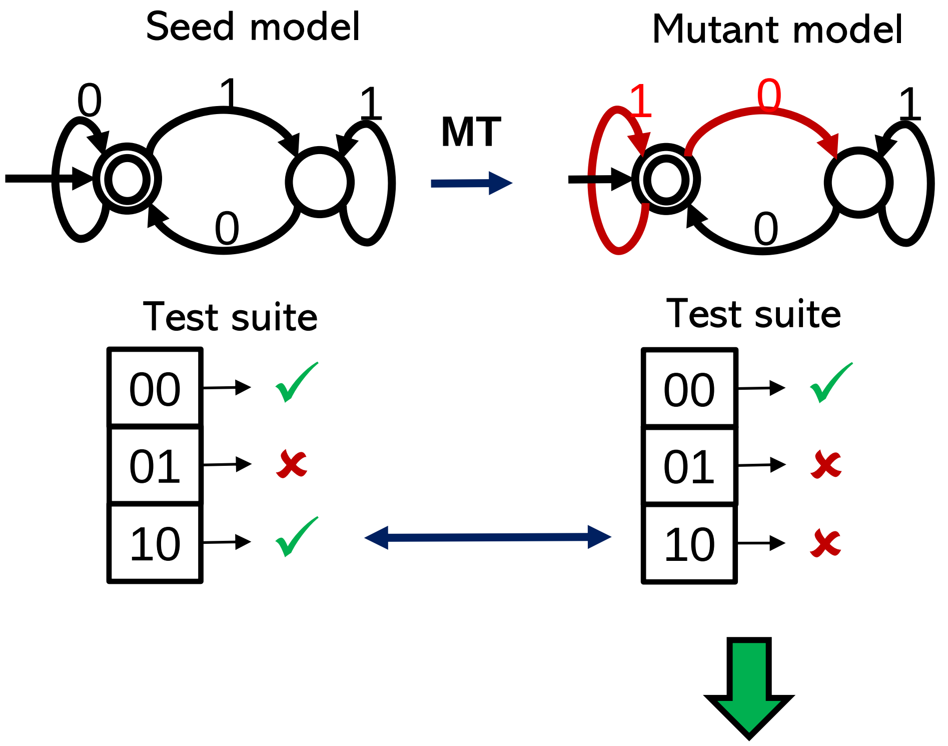
Test suite

00	✓
01	✗
10	✗

Mutation testing for automata

630I

ProxyHands



Test suite

00	✓
01	✗
10	✓

Test suite

00	✓
01	✗
10	✗

The mutant is killed ✓

630I

ProxyHands

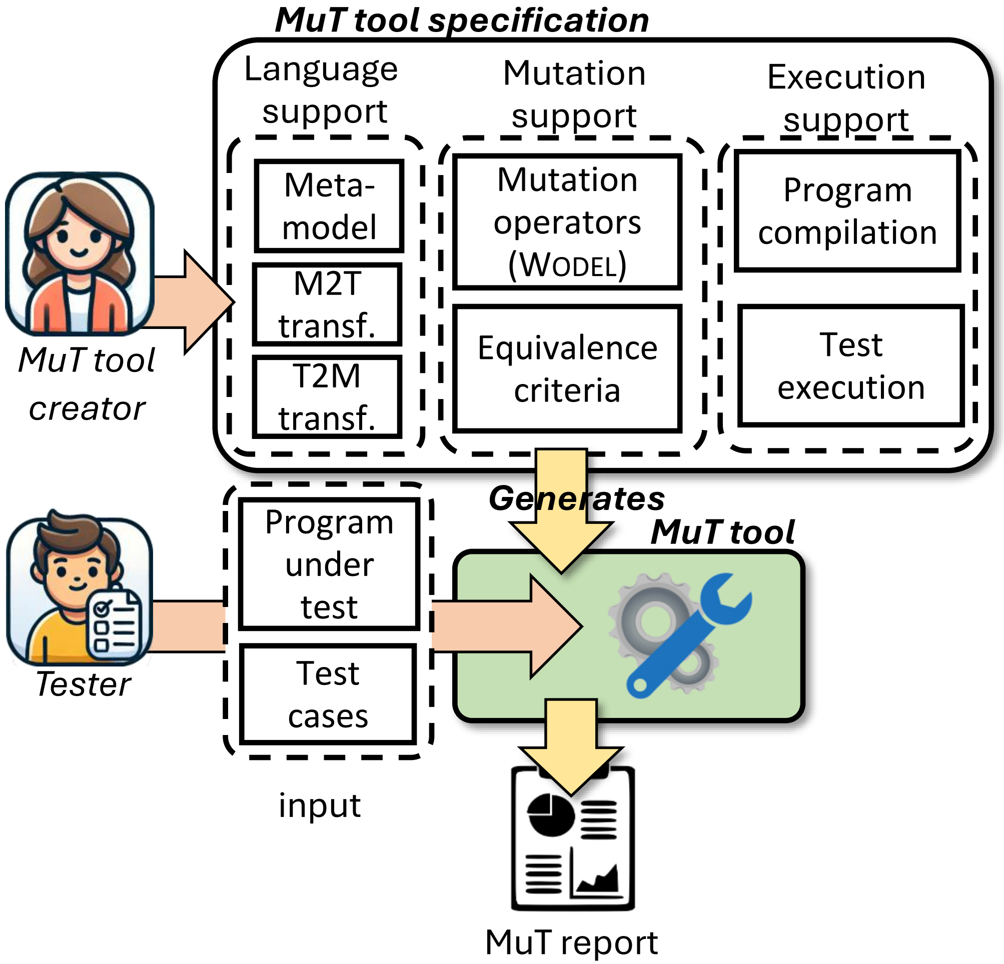
Motivation

- However, the existing MuT tools are
  - Specific for a language
  - Encoded by hand
  - They incur in high-costs of maintenance
- To alleviate such inconveniences, we propose **Wodel-Test**
  - A model-based solution to engineer language-specific MuT tools

Wodel-Test

- A model-based solution to engineer mutation testing tools

- MuT tools for automata, logic circuits, Java, ATL, chatbots, etc.



MuT tool for chatbots

- We have used Wodel-Test to engineer a MuT tool for task-oriented chatbots
- The solution uses the intent-based chatbot meta-model created by S. Pérez-Soler et al. [1]

[1] S. Pérez-Soler, E. Guerra, and J. de Lara. Model-driven chatbot development. In ER, volume 12400 of LNCS, pages 207–222. Springer, 2020

What is a task-oriented chatbot?

- A task-oriented chatbot is a software application used in natural language and designed to solve a specific task
  - e.g., booking a ticket, ordering a pizza, setting a medical appointment

- Via text or speech recognition
- In the recent years, the use of chatbots has increased

630I

ProxyHands

Dialogflow

RASA

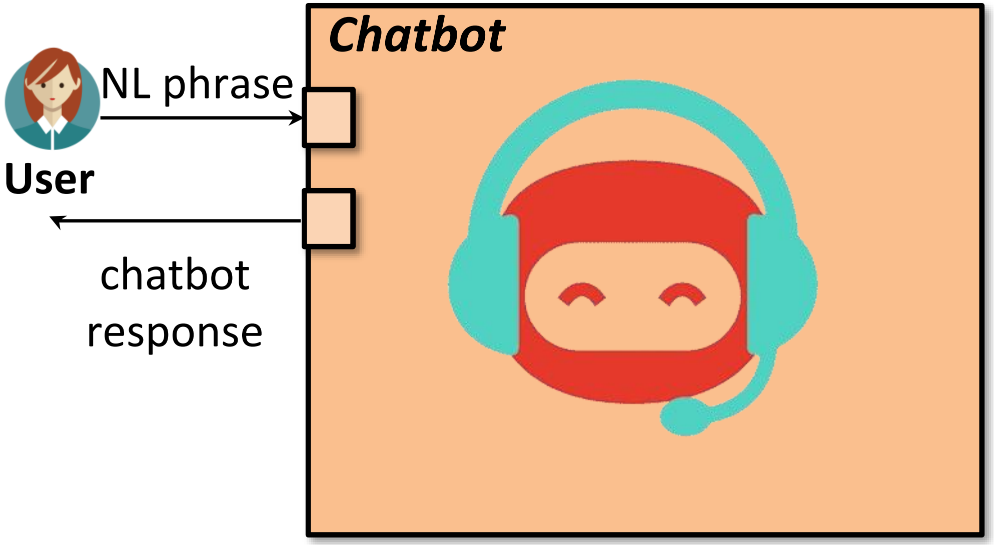
…and many more

- Since 2022, we also have open-domain chatbots (ChatGPT, etc.) which engage in conversations on any topic, and which we do not cover in this work

How do chatbots work?

630I

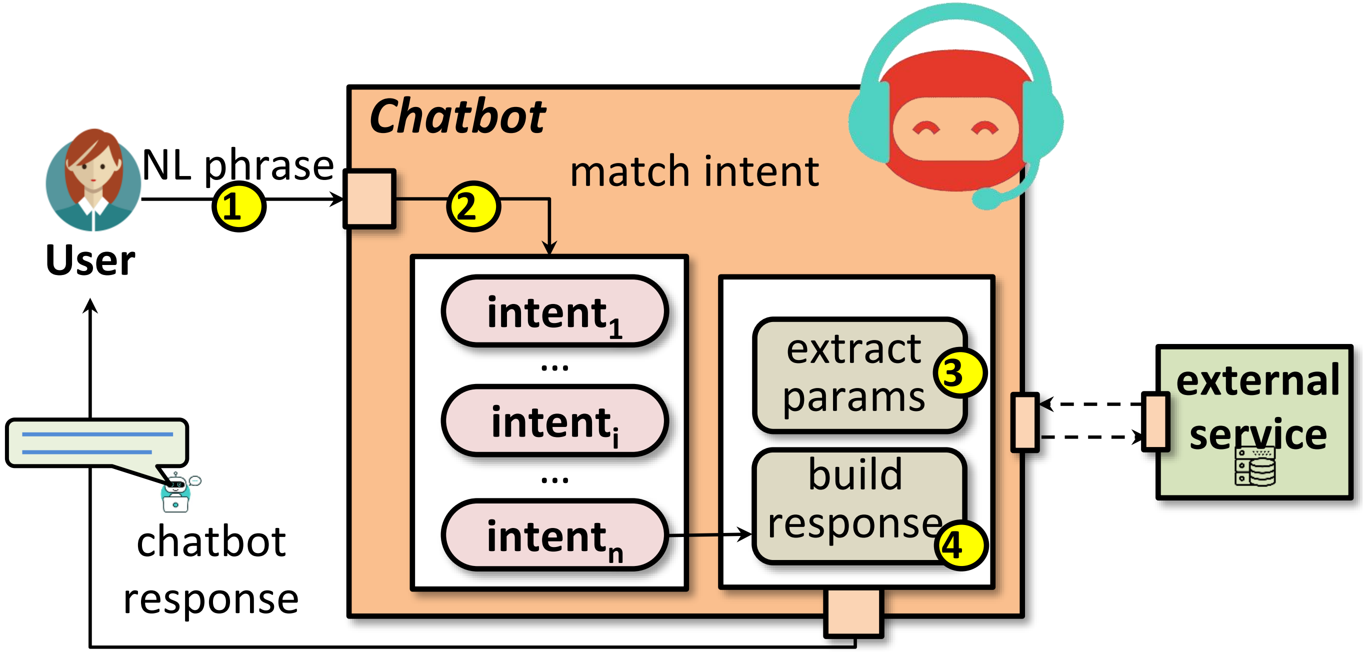
ProxyHands



How do chatbots work?

630I

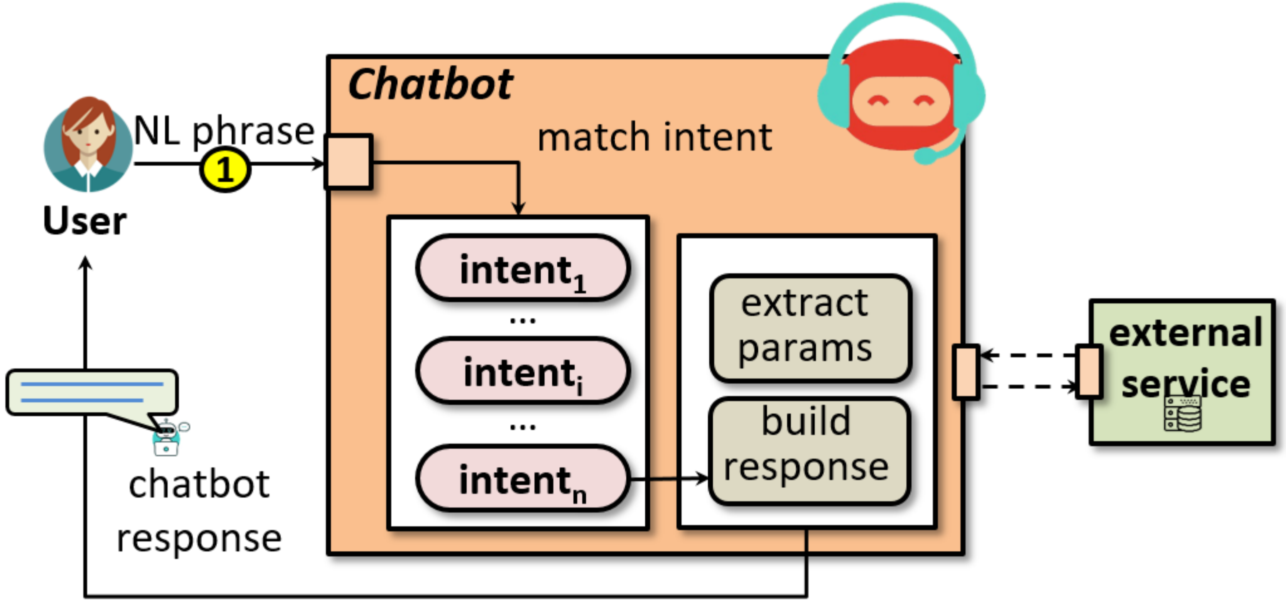
ProxyHands



How do chatbots work?

630I

ProxyHands



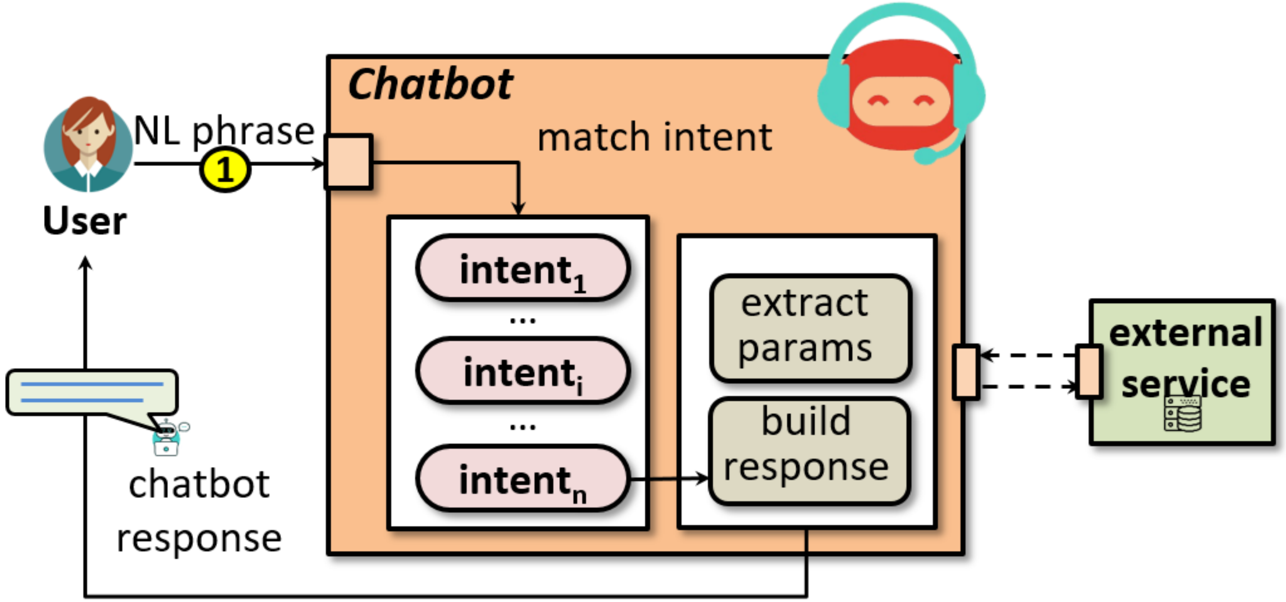
1. The user sends a natural language message to the chatbot **Utterances**

Utterances (user says)
Hi there!
I need to fly from Madrid to Seville on Thursday at 8 AM
Good bye!

How do chatbots work?

630I

ProxyHands

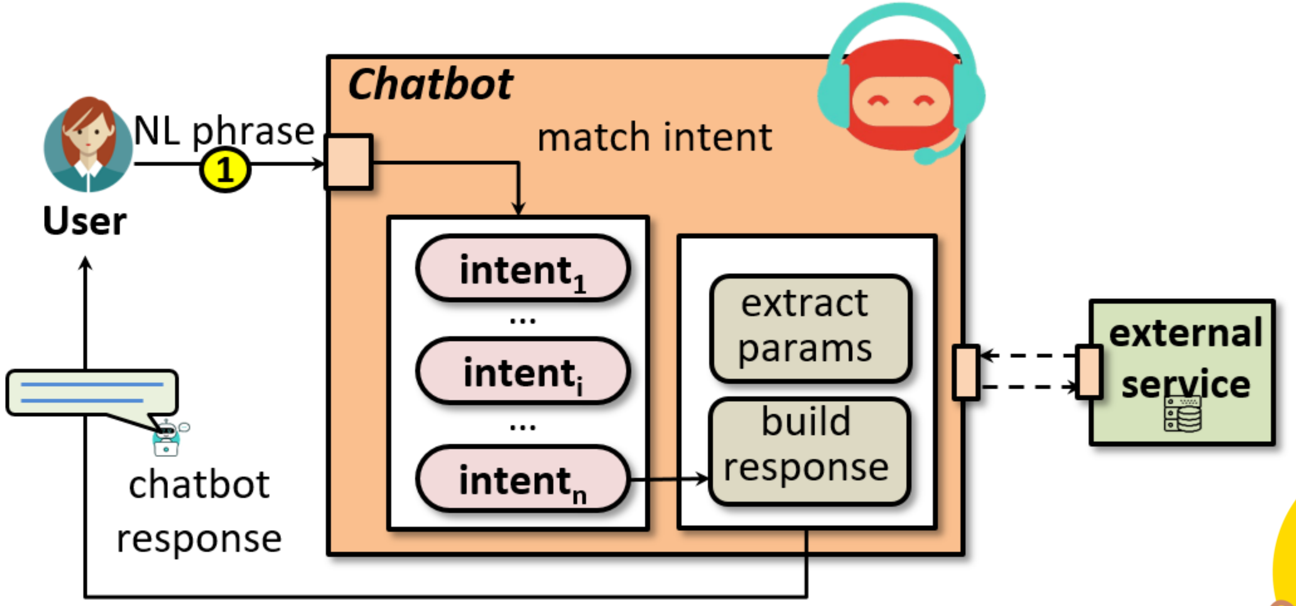


1. The user sends a natural language message to the chatbot
2. The chatbot tries to match the message with an intention

How do chatbots work?

630I

ProxyHands



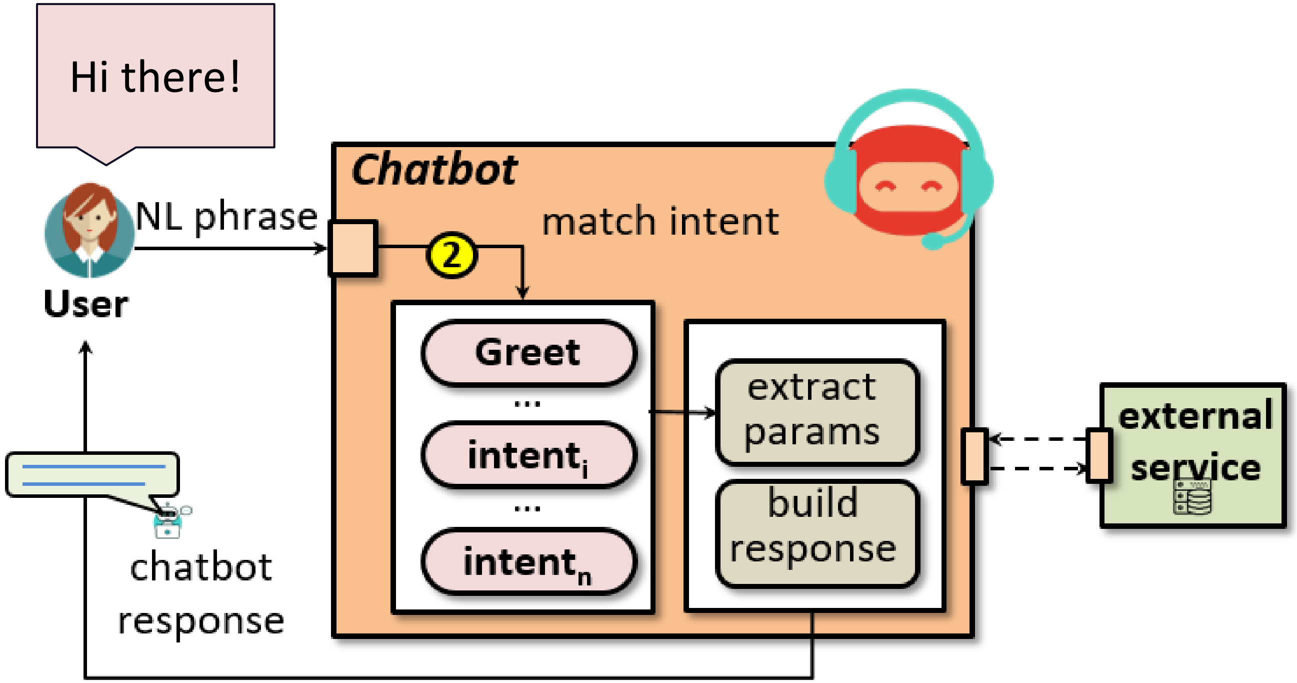
1. The user sends a natural language message to the chatbot
2. The chatbot tries to match the message with an intention

??
Intention?

How do chatbots work?

630I

ProxyHands



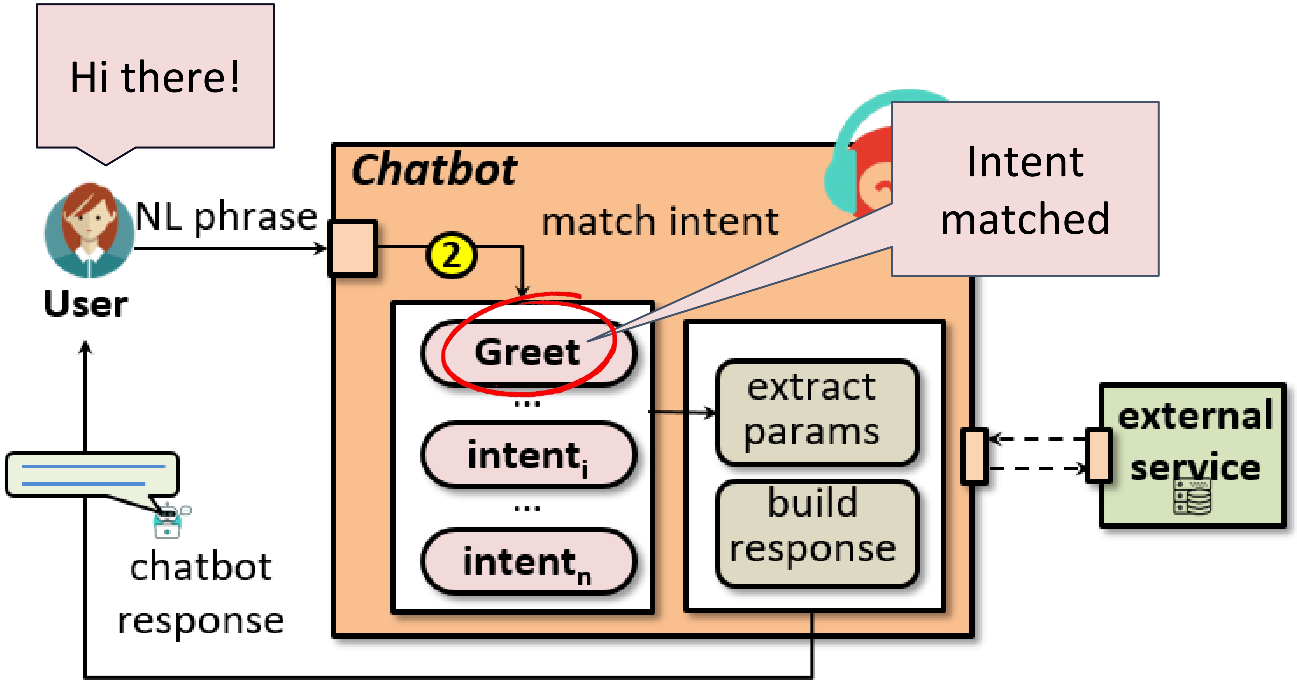
Intent: Match the user interaction with an intention

User says	Intent
Hi there!	

How do chatbots work?

630I

ProxyHands



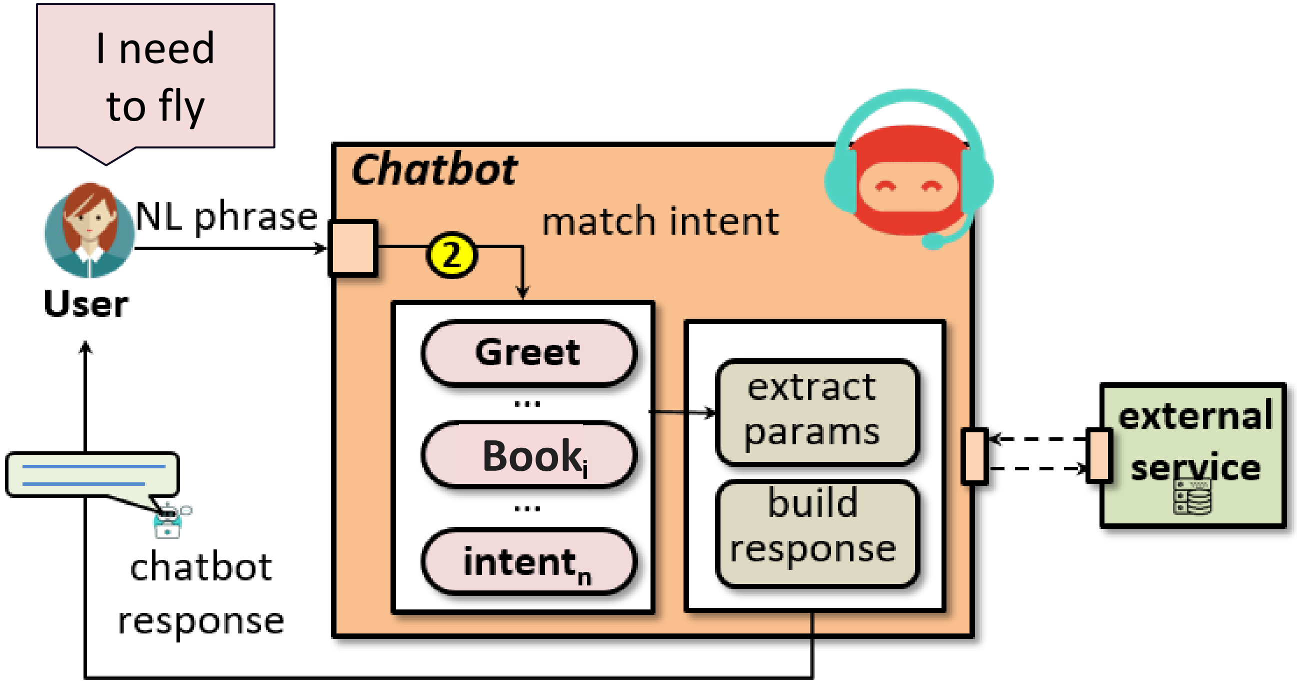
Intent: Match the user interaction with an intention

User says	Intent
Hi there!	Greet

630I

ProxyHands

How do chatbots work?



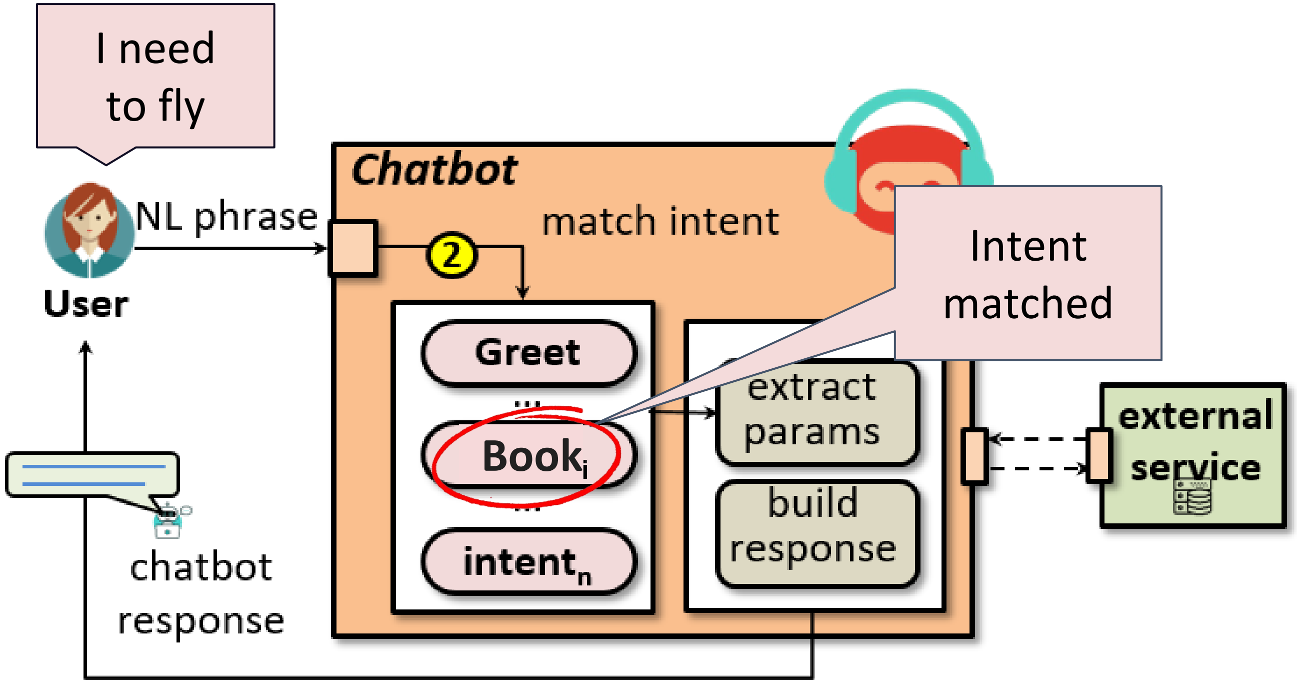
Intent: Match the user interaction with an intention

User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	

How do chatbots work?

630I

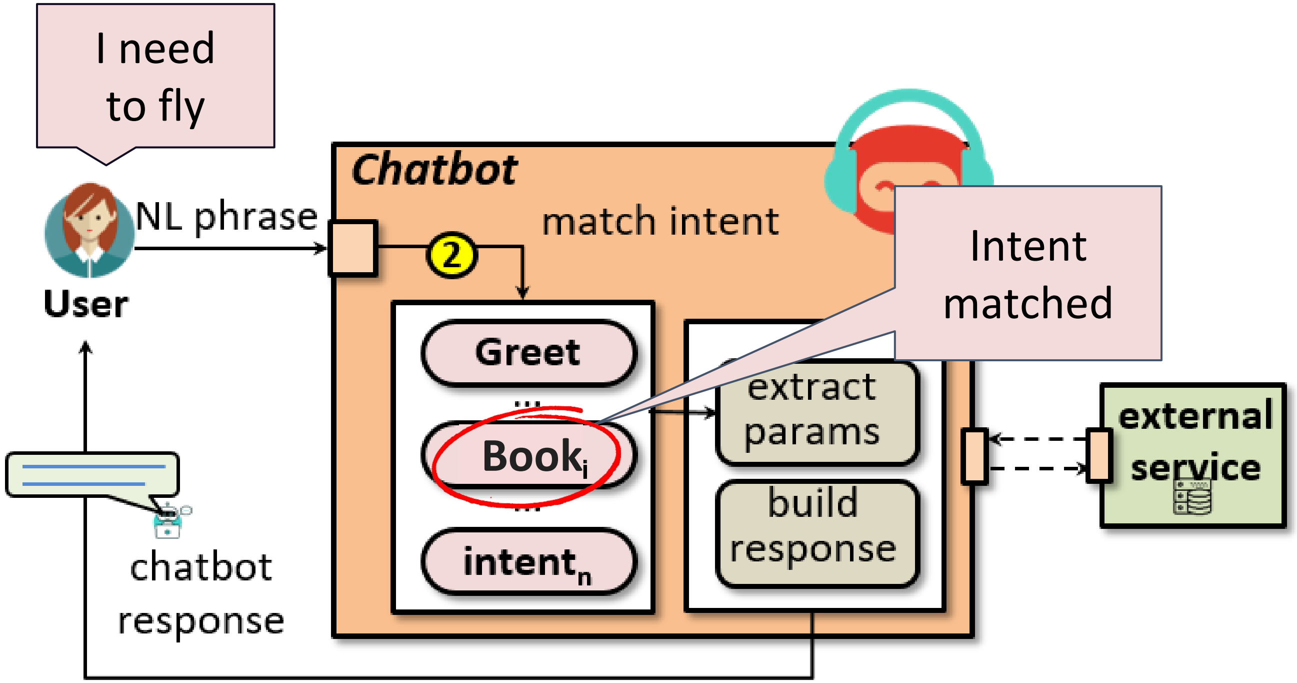
ProxyHands



Intent: Match the user interaction with an intention

User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	Book a flight

How do chatbots work?



HOW?!

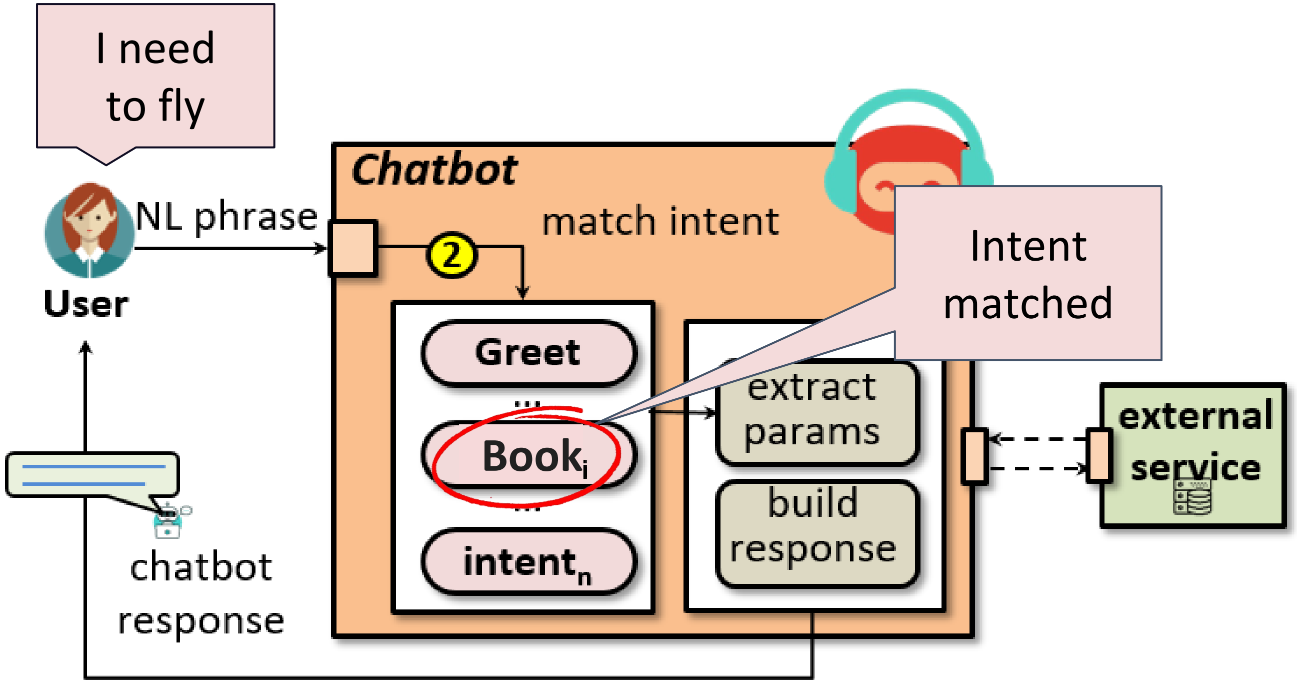
Intent: Match the user interaction with an intention

User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	Book a flight

How do chatbots work?

630I

ProxyHands



Intent: Match the user interaction with an intention

User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	Book a flight

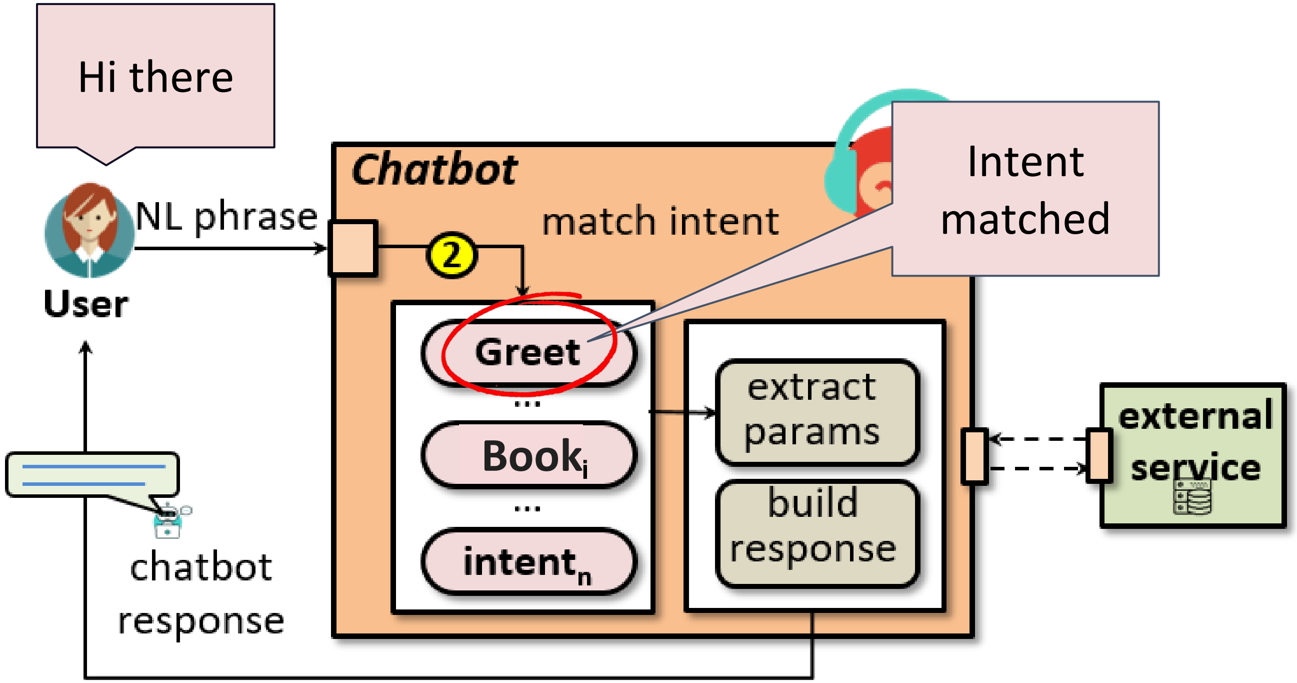
HOW?!

Providing **training phrases:** a set of examples that users can use to express an intention. Required for matching inputs with intents

How do chatbots work?

630I

ProxyHands



Training phrases: a set of examples that users can use to express an intention

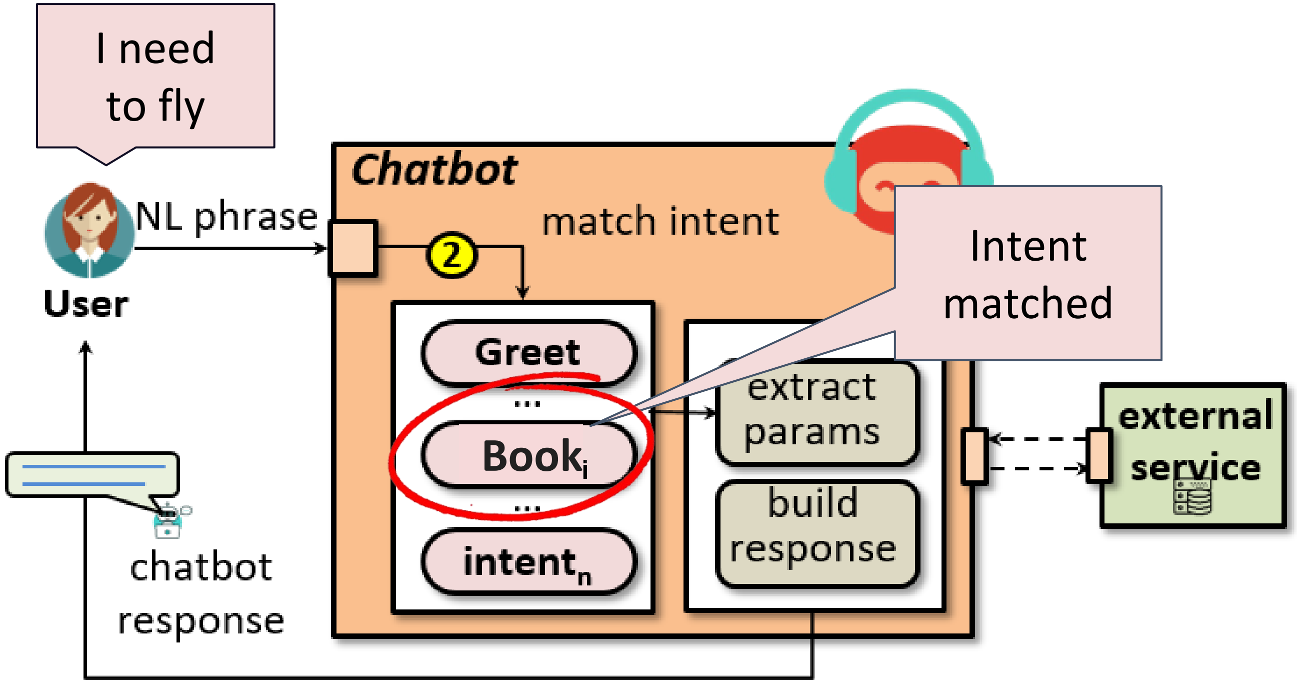
- Must be provided with the intent

Training phrase	Intent
Hi there!	Greet
Hello	Greet
Hi	Greet
Hey	Greet

How do chatbots work?

630I

ProxyHands



Training phrases: a set of examples that users can use to express an intention

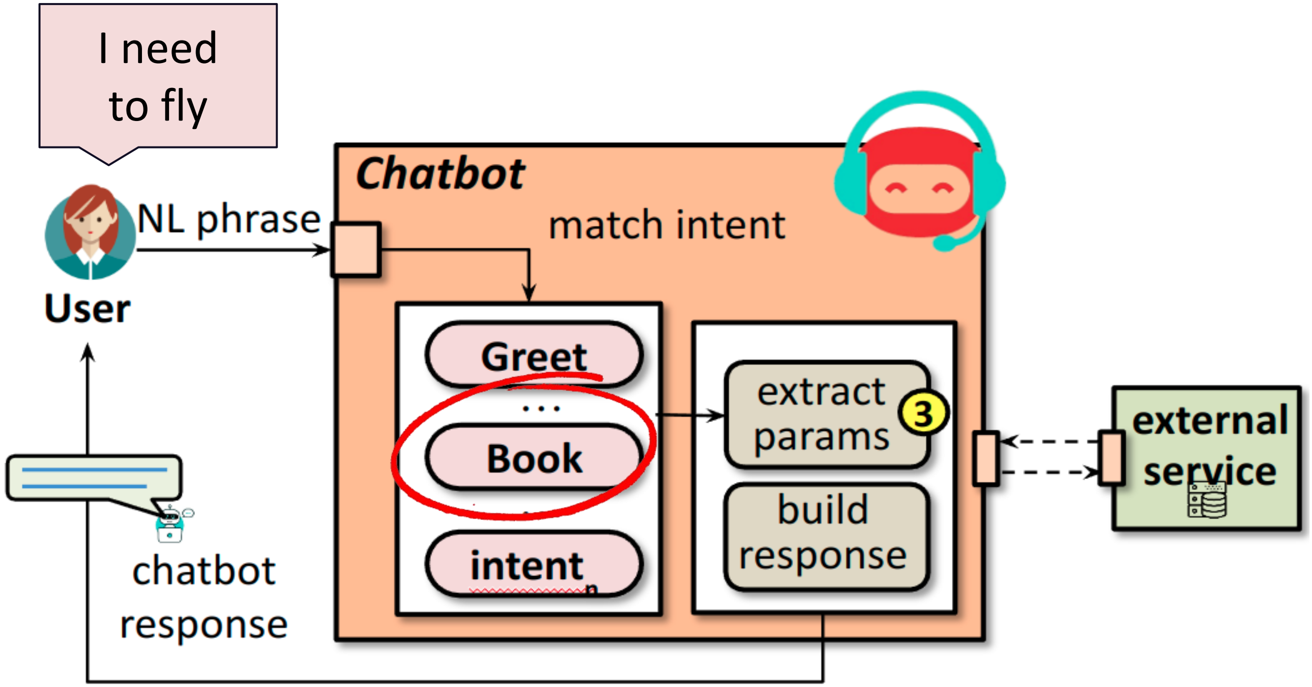
- Must be provided with the intent

Training phrase	Intent
Airplane ticket from Madrid to Barcelona tomorrow at 10 AM	Book a flight
Flight from Madrid to Bilbao on 19/10/2024 at 11:30	Book a flight

How do chatbots work?

630I

ProxyHands



3. Chatbot extracts information from the message or asks for missing information

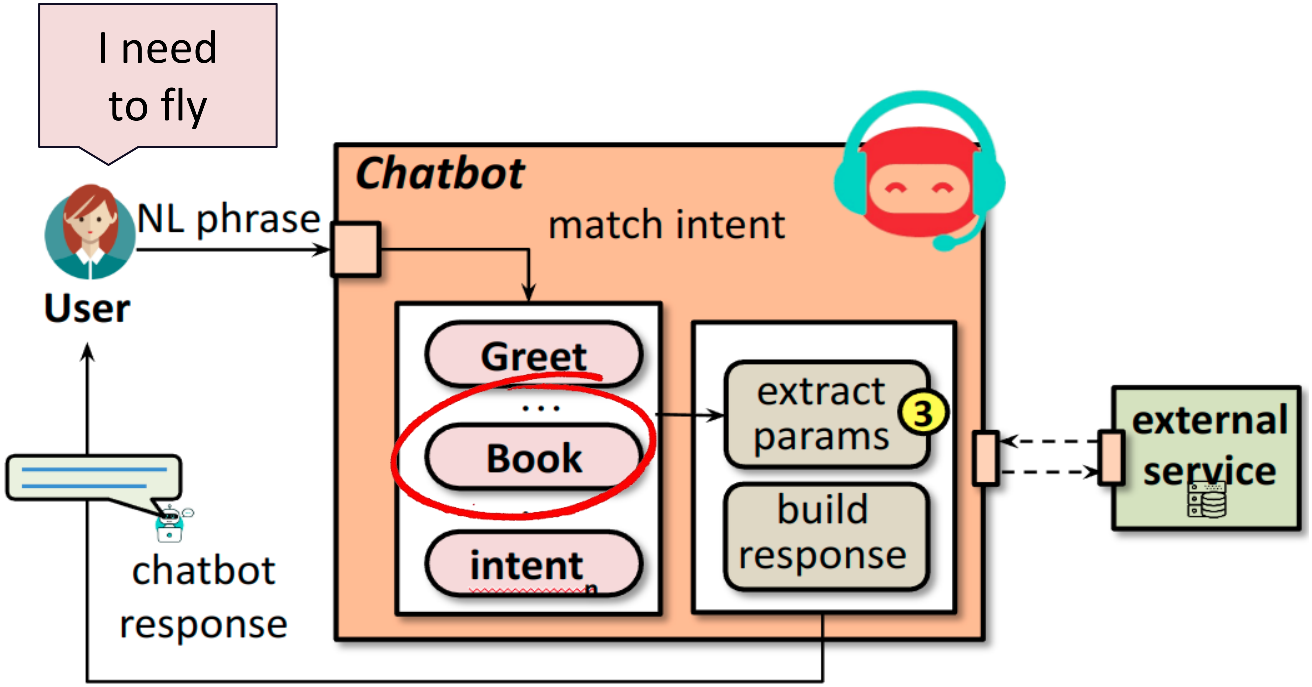
User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	Book a flight

How do chatbots work?

630I

ProxyHands

3. Chatbot extracts information from the message or asks for missing information



User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	Book a flight

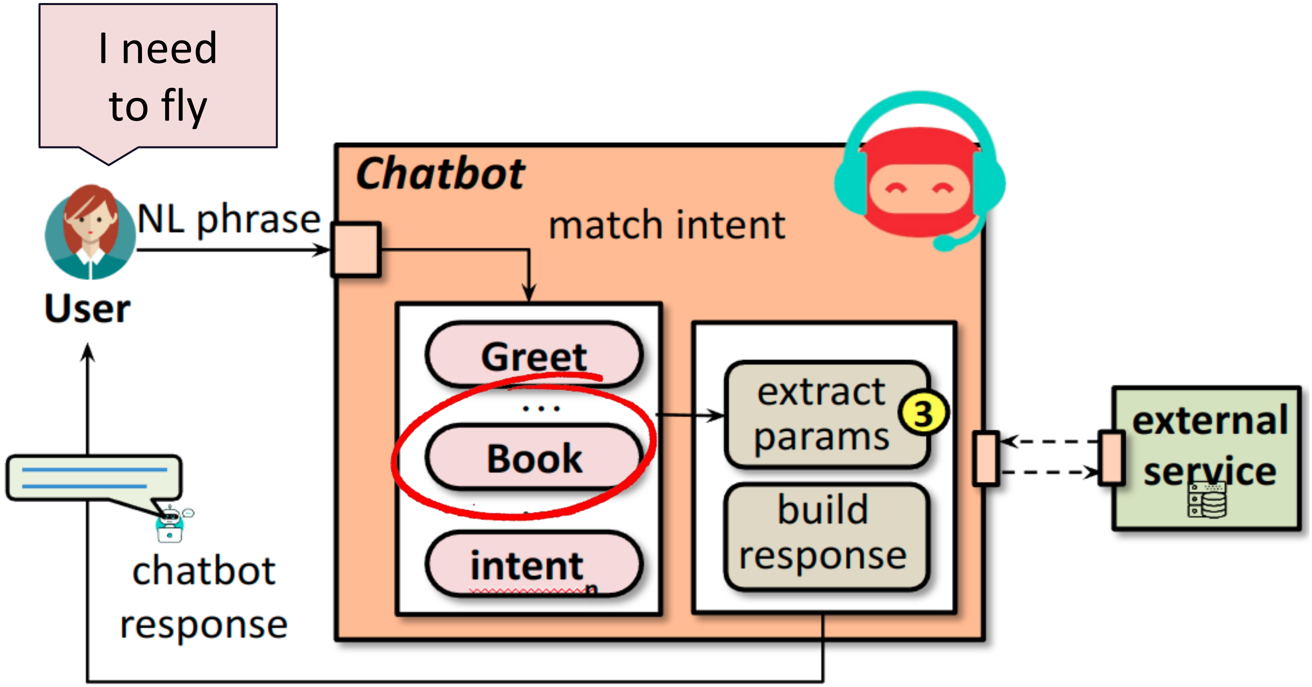
At this point, the chatbot extracts key information from the input: **parameters**

From:Madrid to:Seville when:Thu. At 8 AM

How do chatbots work?

630I

ProxyHands



3. Chatbot extracts information from the message or asks for missing information

User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	Book a flight

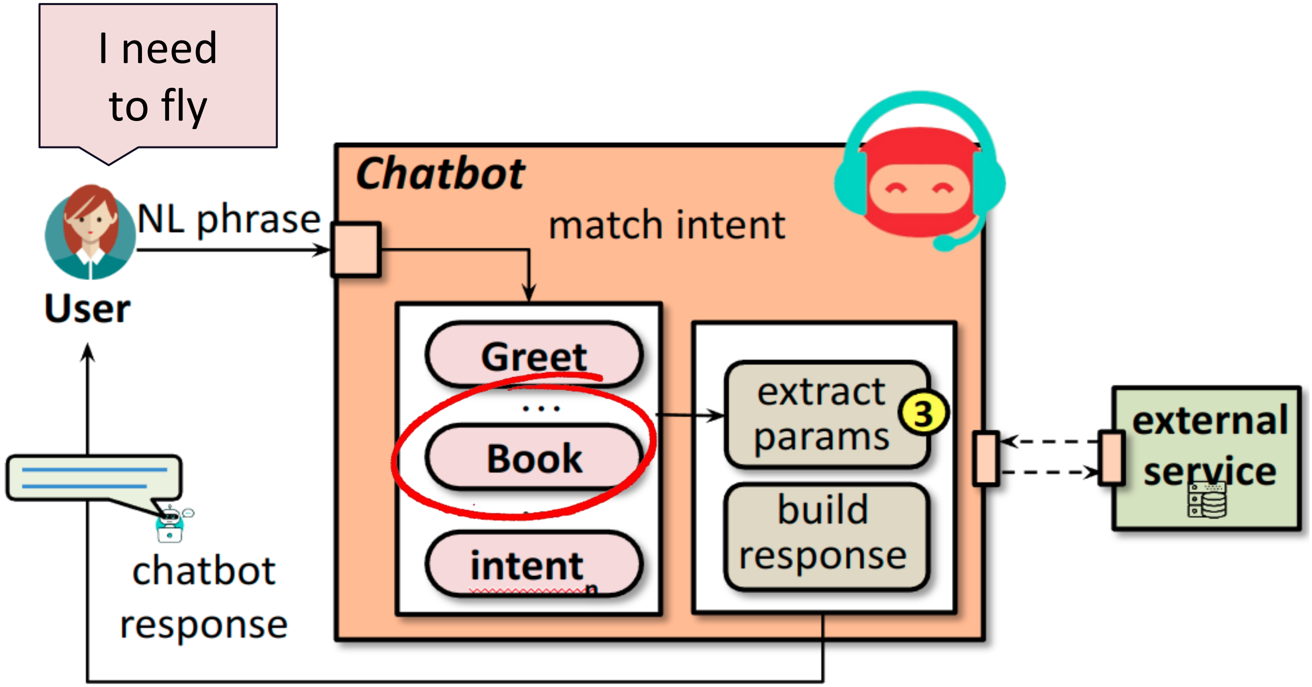
At this point, the chatbot extracts key information from the input: **parameters**

From:Madrid to:Seville when:Thu. At 8 AM

City

630I

How do chatbots work?



3. Chatbot extracts information from the message or asks for missing information

User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	Book a flight

At this point, the chatbot extracts key information from the input: **parameters**

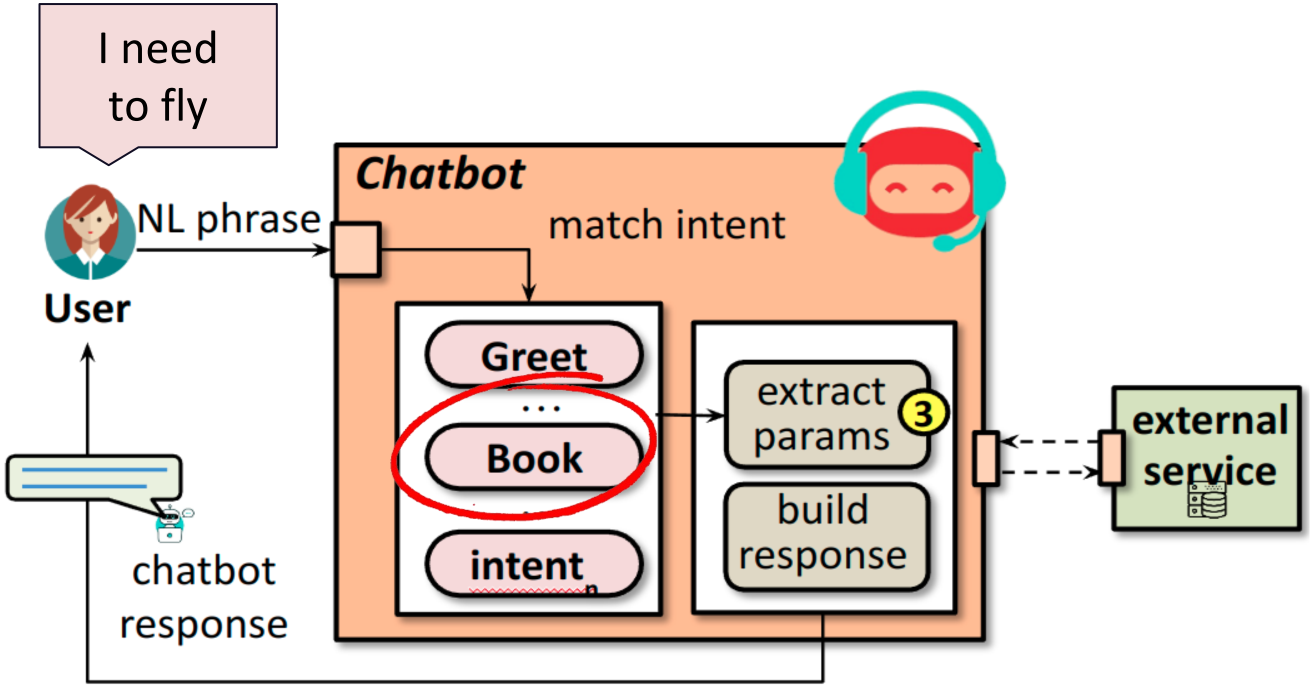
From Madrid to Seville when:Thu. At 8 AM

City entities

How do chatbots work?

630I

ProxyHands



3. Chatbot extracts information from the message or asks for missing information

User says	Intent
I need to fly from Madrid to Seville on Thursday at 8 AM	Book a flight

At this point, the chatbot extracts key information from the input: **parameters**

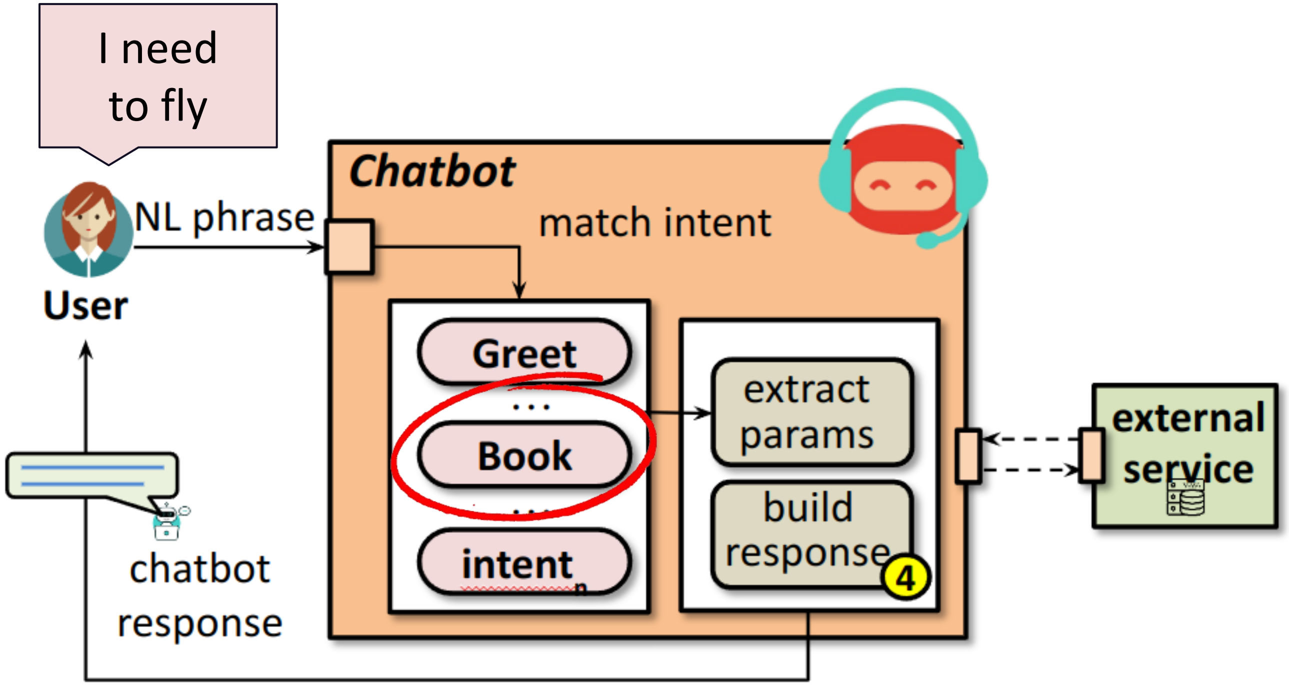
From Madrid to Seville when Thu, At 8 AM

City entities Time

How do chatbots work?

630I

ProxyHands



4. Build the response and send back the response to the user

- Responses to the user:
  - text, images
- External service queries
  - External API rest
  - Database, etc.

User says	Action
I need to fly from Madrid to Seville on Thursday at 8 AM	The price of the ticket is 120$. Provide a card nº and billing name

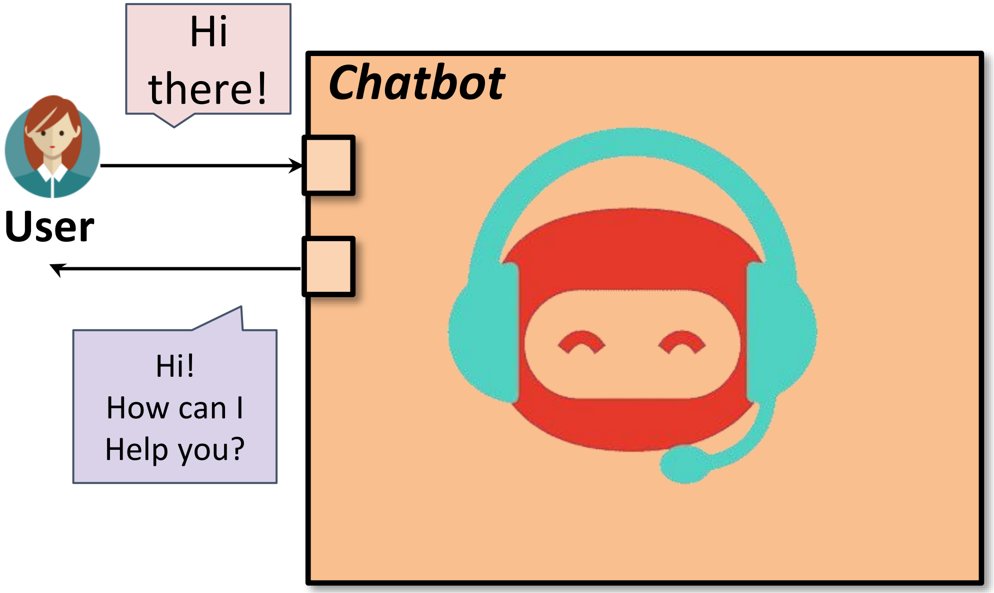
Both, user responses and external services queries: **actions**

City entities Time

Testing chatbots

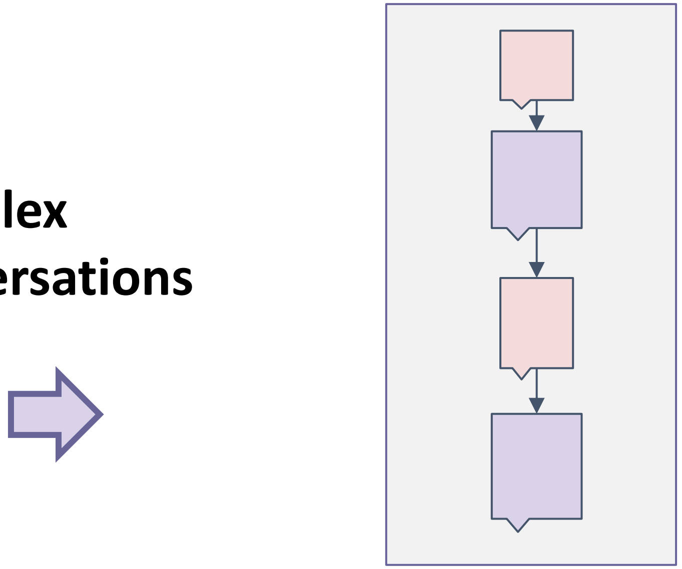
630I

ProxyHands



Testcase input	Testcase output
Hi there!	Hi! How can I help you?

…
complex conversations

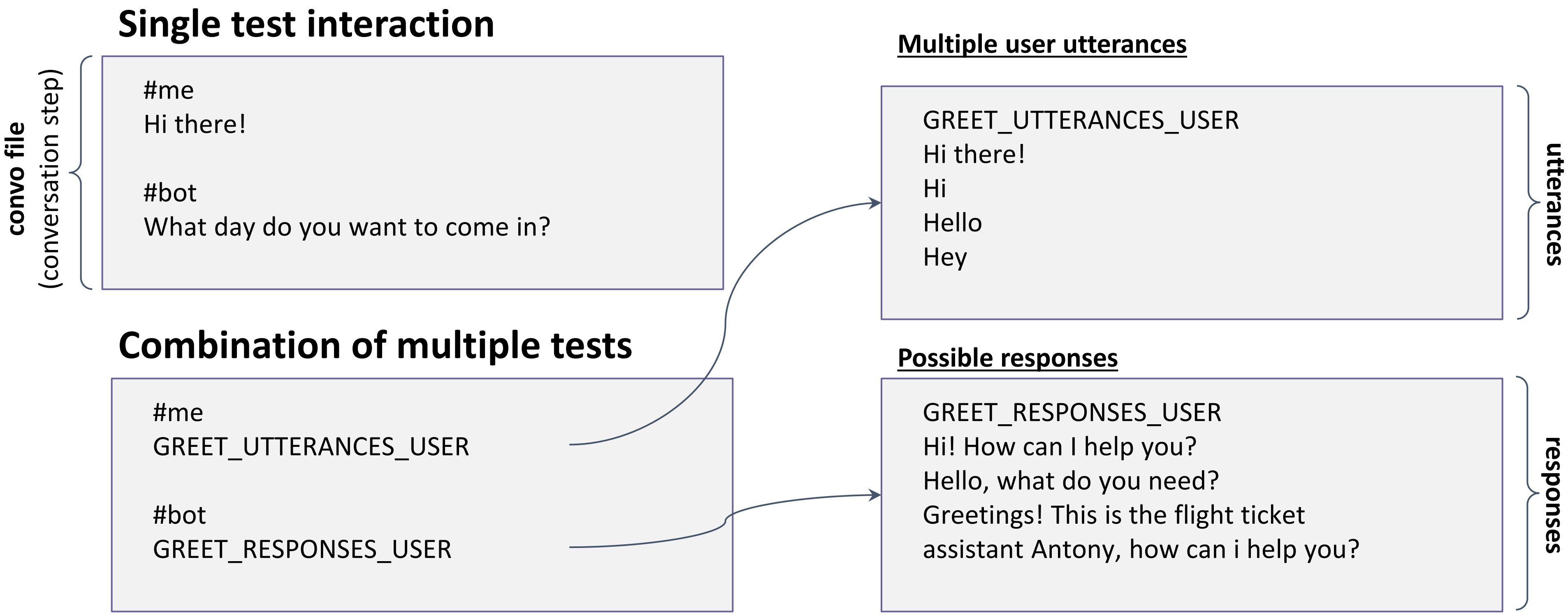


630I

ProxyHands

Testing chatbots

We use **Botium** and Rasa-test as the test suites to test the chatbots



Single test interaction

```
#me
Hi there!

#bot
What day do you want to come in?
```

Combination of multiple tests

```
#me
GREET_UTTERANCES_USER

#bot
GREET_RESPONSES_USER
```

Multiple user utterances

```
GREET_UTTERANCES_USER
Hi there!
Hi
Hello
Hey
```

Possible responses

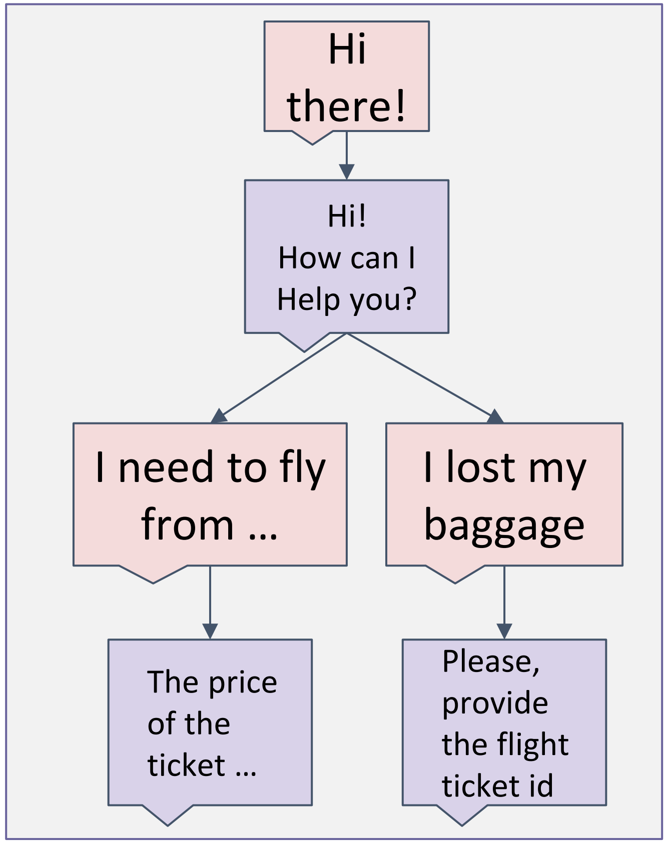
```
GREET_RESPONSES_USER
Hi! How can I help you?
Hello, what do you need?
Greetings! This is the flight ticket
assistant Antony, how can i help you?
```

Testing chatbots

630I

ProxyHands

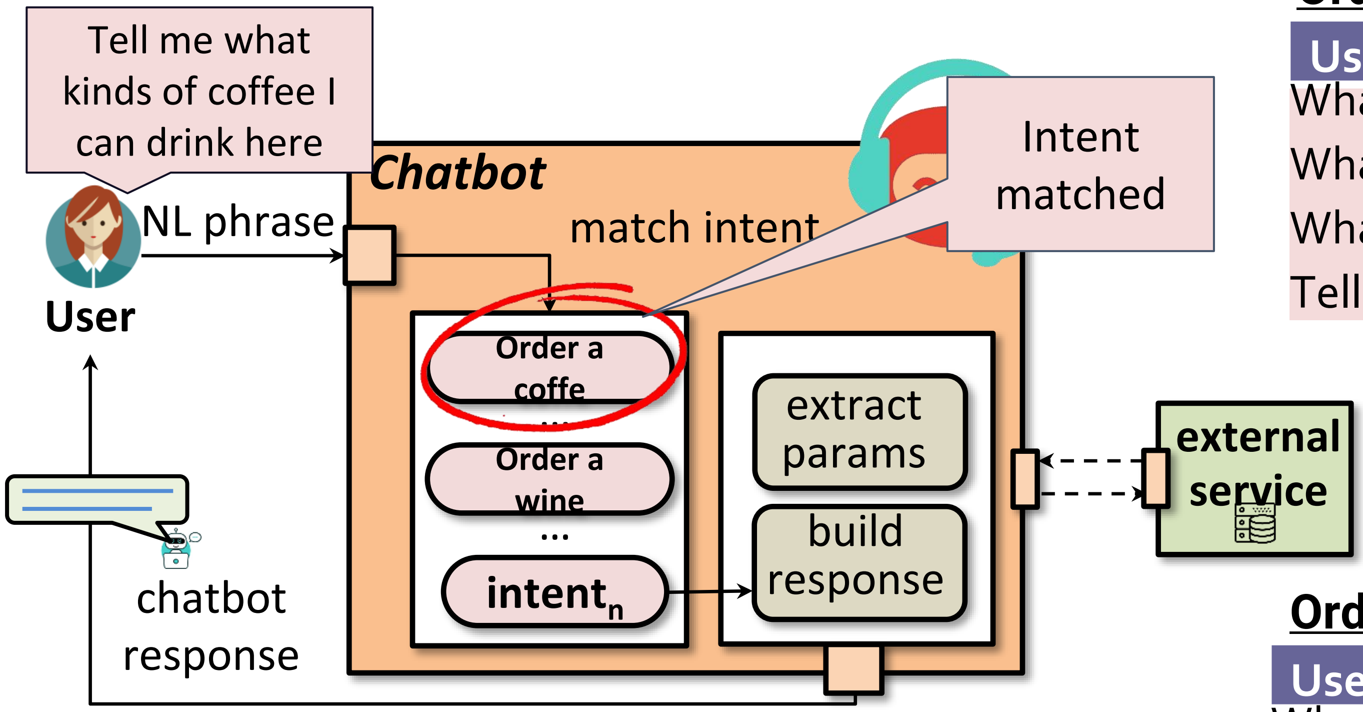
… and complex conversations



Mutation testing for chatbots

630I

ProxyHands



Order a coffee

User says	Action
What kinds of coffee are available?	You can take an expresso or an americano
What kinds of coffee can I order?	
What can I drink here?	
Tell me what drinks there are	

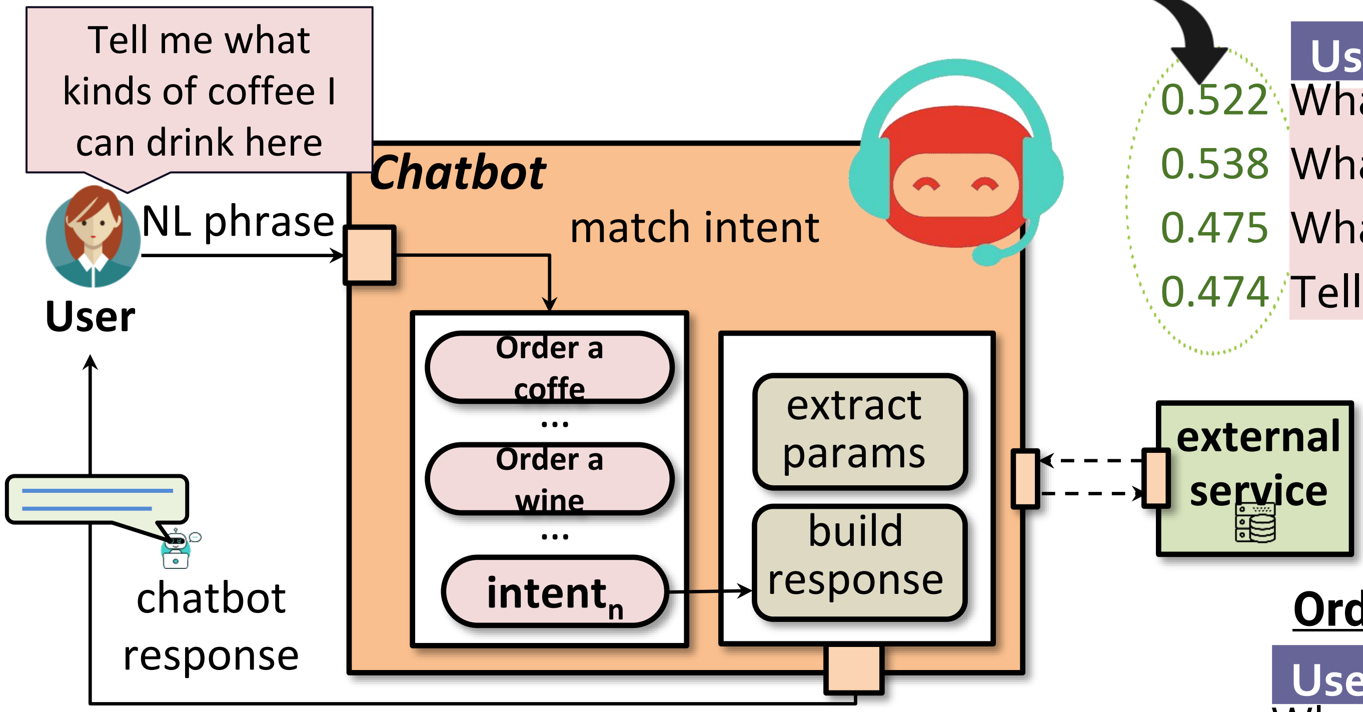
Order a wine

User says	Action
What kinds of wine are available?	You can take a Spanish wine or a French wine
What kinds of wine can I order?	
What can I drink here?	
Tell me what drinks there are	

Mutation testing for chatbots

630I

ProxyHands



Semantic similarity

Order a coffee: Keeps the two most different phrases

	User says	Action
0.522	What kinds of coffee are available?	You can take an expresso or an americano
0.538	What kinds of coffee can I order?	
0.475	What can I drink here?	
0.474	Tell me what drinks there are	

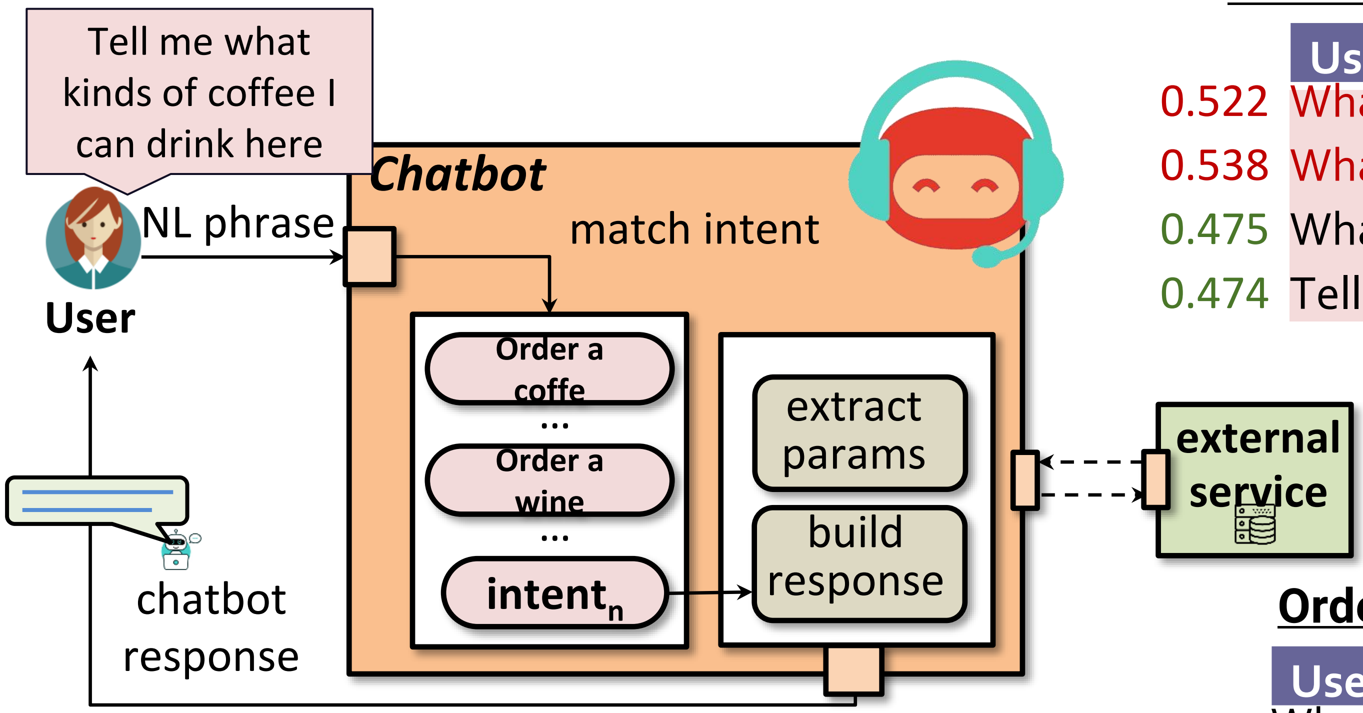
Order a wine

User says	Action
What kinds of wine are available?	You can take a Spanish wine or a French wine
What kinds of wine can I order?	
What can I drink here?	
Tell me what drinks there are	

Mutation testing for chatbots

630I

ProxyHands



Order a coffee: Keeps the two most different phrases

	User says	Action
0.522	What kinds of coffee are available?	You can take an expresso or an americano
0.538	What kinds of coffee can I order?	
0.475	What can I drink here?	
0.474	Tell me what drinks there are	

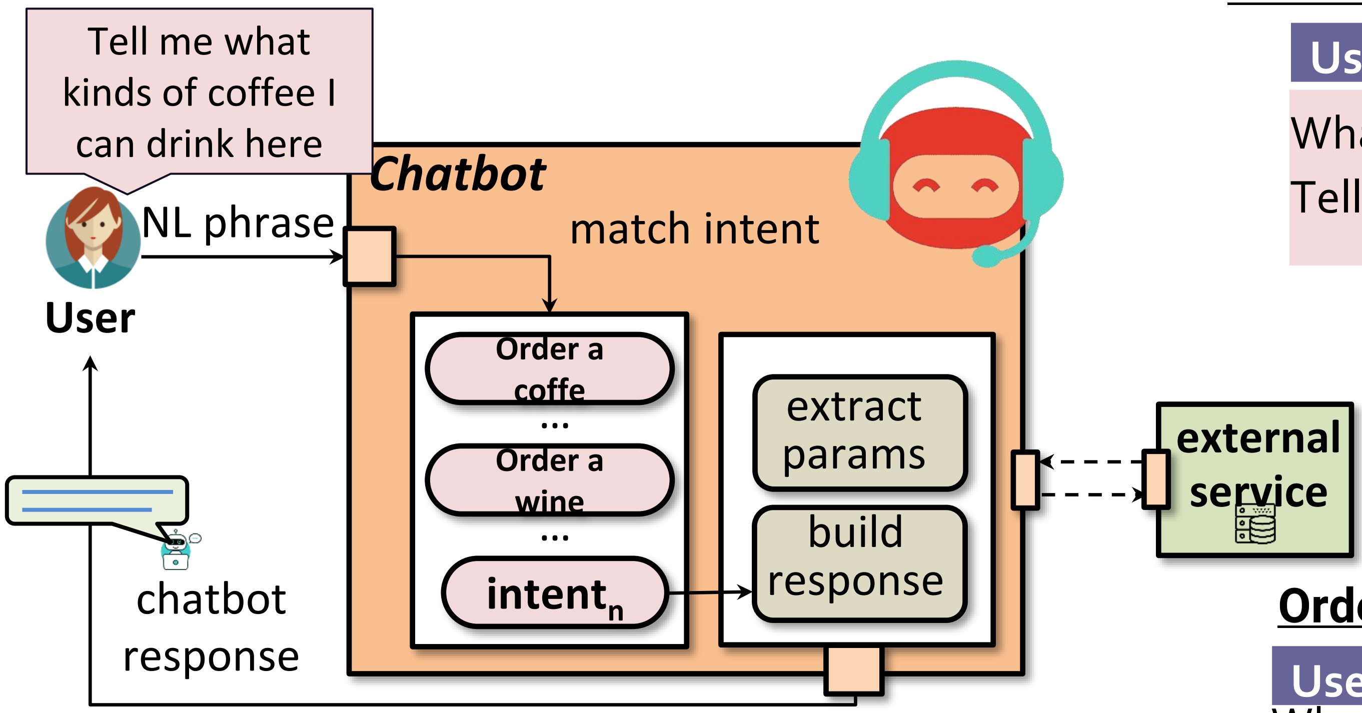
Order a wine

User says	Action
What kinds of wine are available?	You can take a Spanish wine or a French wine
What kinds of wine can I order?	
What can I drink here?	
Tell me what drinks there are	

Mutation testing for chatbots

630I

ProxyHands



Order a coffee

User says	Action
What can I drink here? Tell me what drinks there are	You can take an expresso or an americano

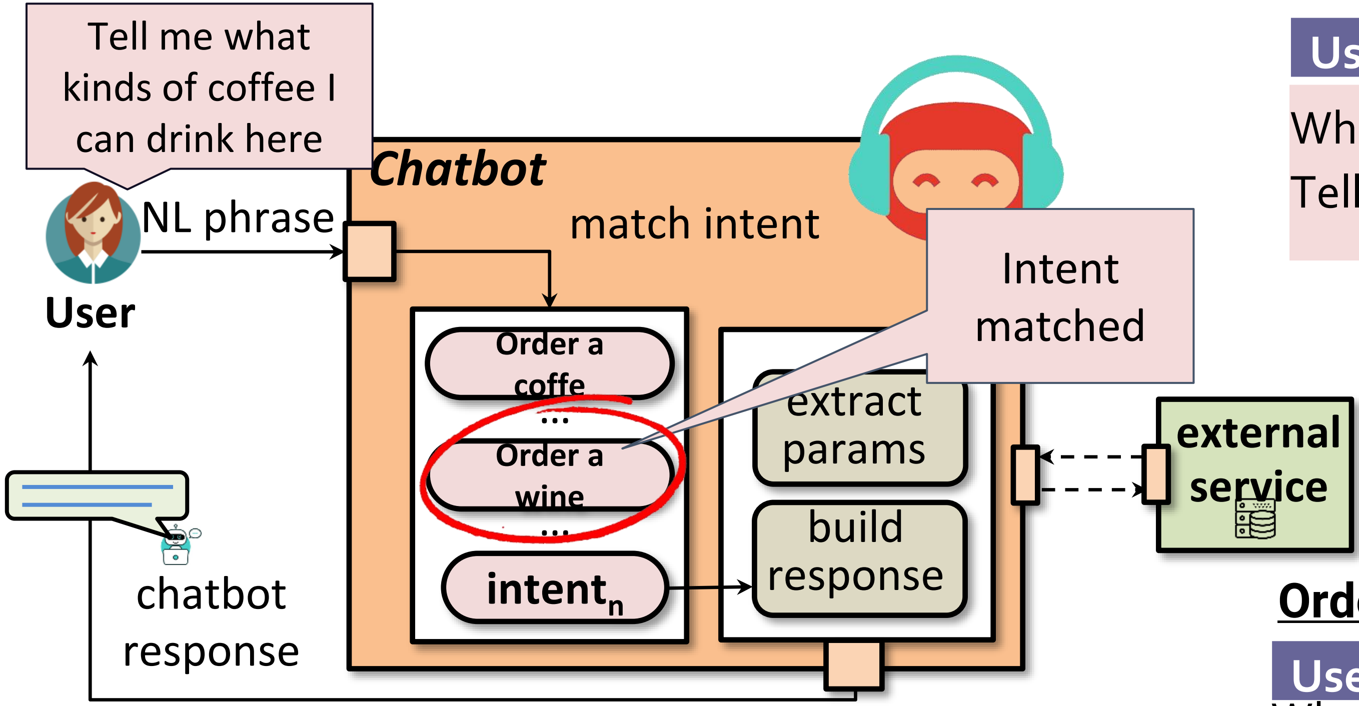
Order a wine

User says	Action
What kinds of wine are available? What kinds of wine can I order? What can I drink here? Tell me what drinks there are	You can take a Spanish wine or a French wine

Mutation testing for chatbots

630I

ProxyHands



Order a coffee

User says	Action
What can I drink here? Tell me what drinks there are	You can take an expresso or an americano

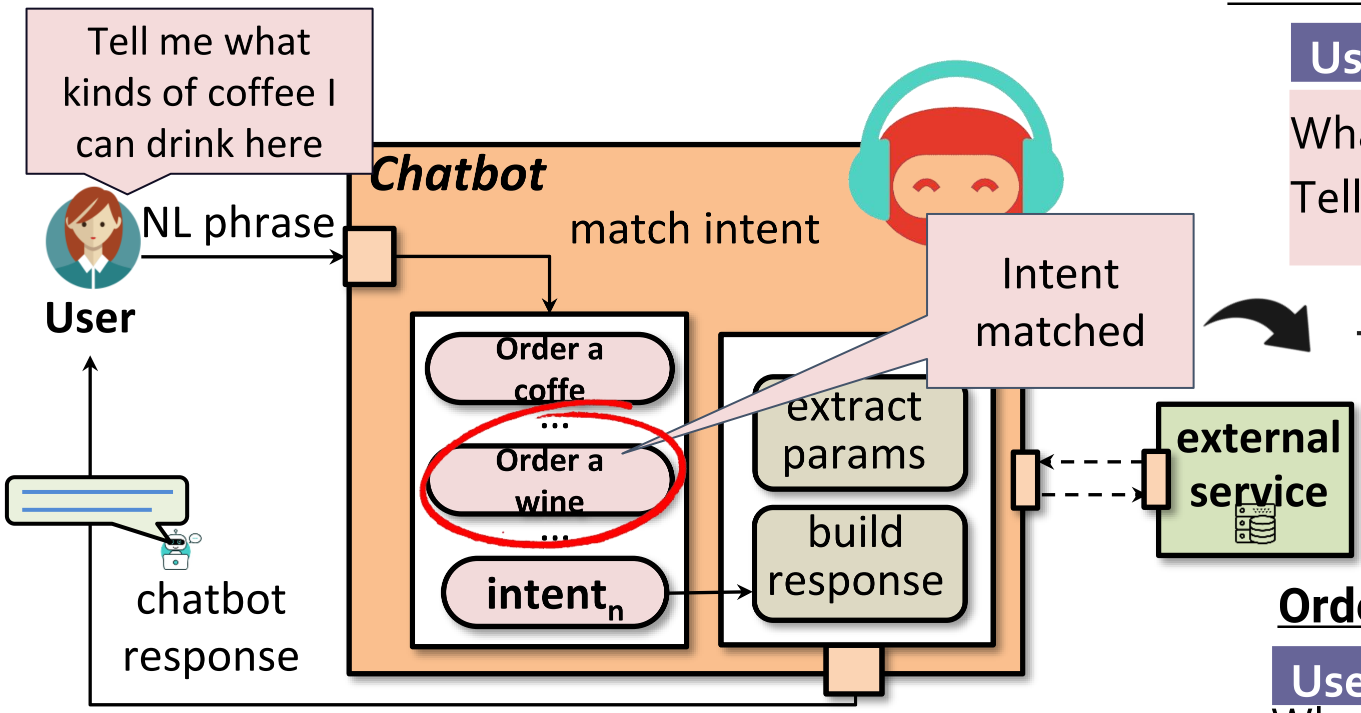
Order a wine

User says	Action
What kinds of wine are available? What kinds of wine can I order? What can I drink here? Tell me what drinks there are	You can take a Spanish wine or a French wine

Mutation testing for chatbots

630I

ProxyHands



Order a coffee

User says	Action
What can I drink here?	You can take an expresso or an americano
Tell me what drinks there are	

Test-suite ✗

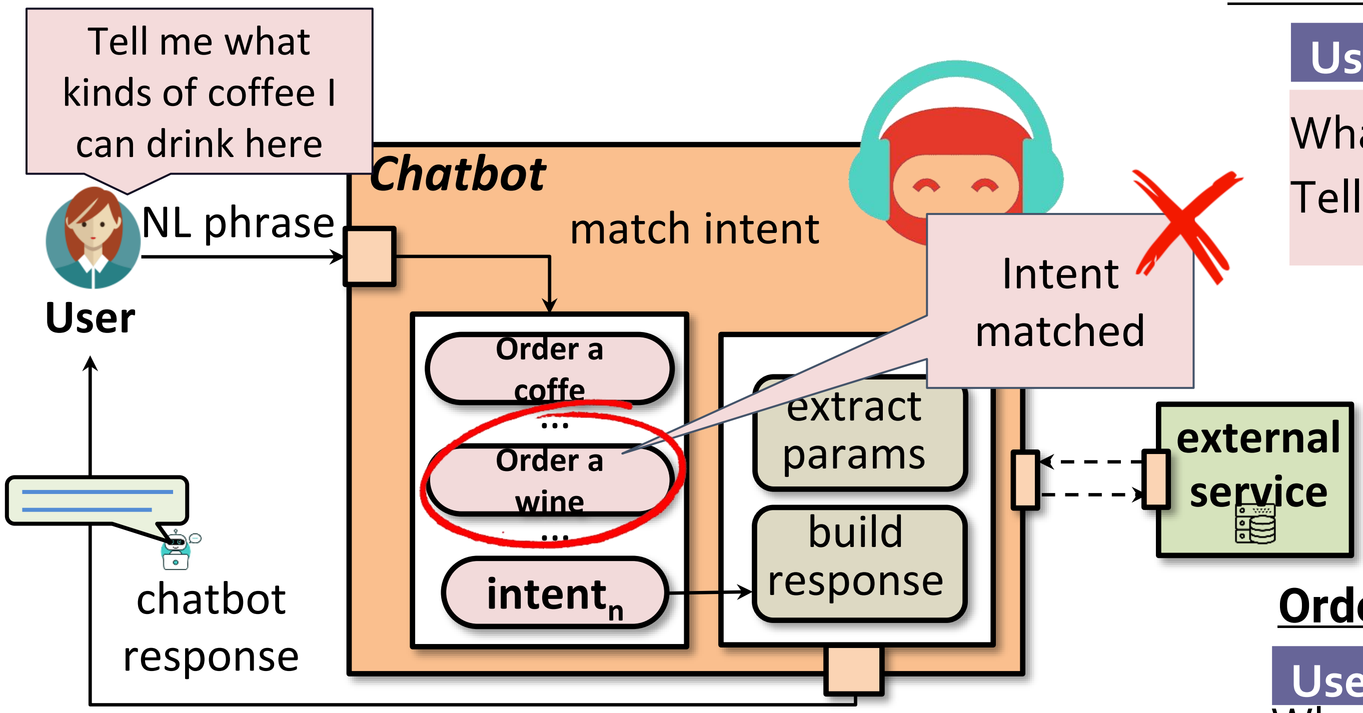
Order a wine

User says	Action
What kinds of wine are available?	You can take a Spanish wine or a French wine
What kinds of wine can I order?	
What can I drink here?	
Tell me what drinks there are	

Mutation testing for chatbots

630I

ProxyHands



Order a coffee

User says	Action
What can I drink here? Tell me what drinks there are	You can take an expresso or an americano

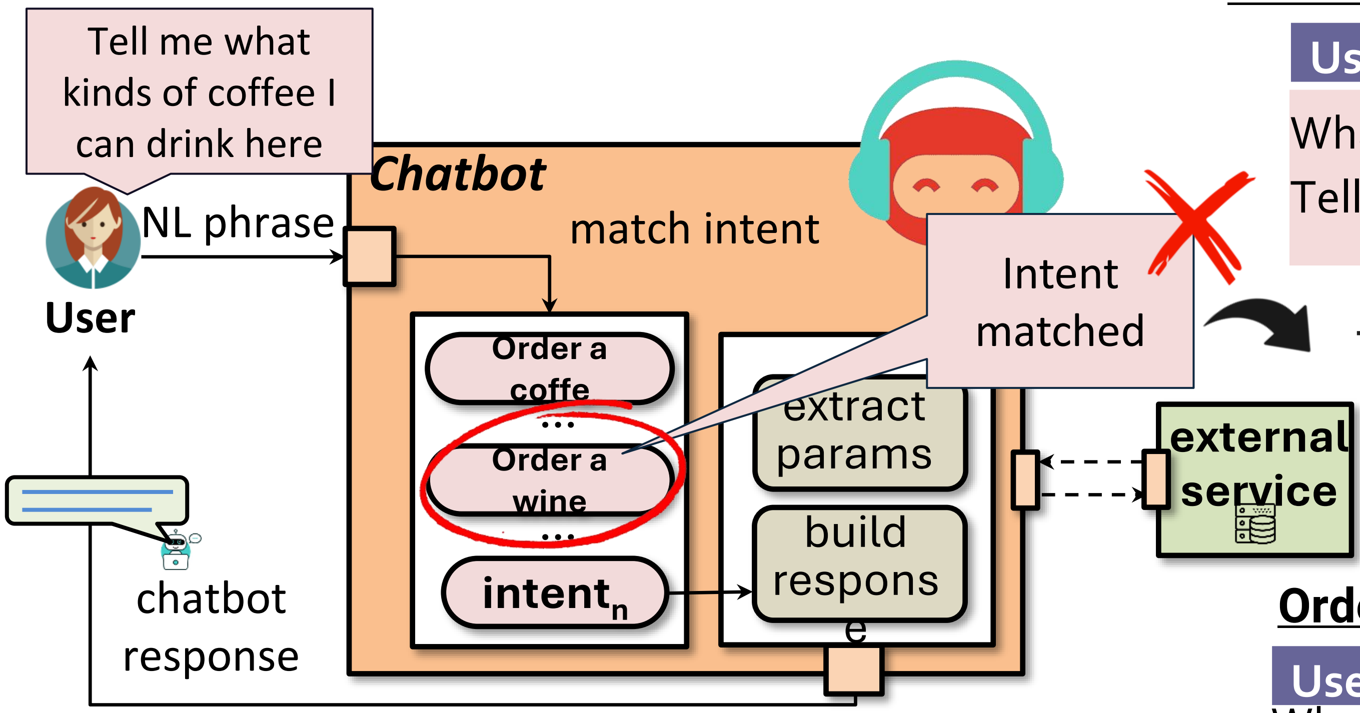
Order a wine

User says	Action
What kinds of wine are available? What kinds of wine can I order? What can I drink here? Tell me what drinks there are	You can take a Spanish wine or a French wine

Mutation testing for chatbots

630I

ProxyHands



Order a coffee

User says	Action
What can I drink here? Tell me what drinks there are	You can take an expresso or an americano

Test-suite ✓

Order a wine

User says	Action
What kinds of wine are available? What kinds of wine can I order? What can I drink here? Tell me what drinks there are	You can take a Spanish wine or a French wine

630I

ProxyHands

Mutation operators for chatbots

Operators for training phrases	
DP_{max}	Deletes the most representative phrase of an intent
DP_{min}	Deletes the most different phrase of an intent
DPWP	Deletes training phrases with required parameter
DPWL	Deletes training phrases with literal
$K2P_{max}$	Keeps the 2 most representative phrases
$K2P_{min}$	Keeps the 2 most different phrases
MP_{max}	Moves the most representative phrase to the most similar intent
MP_{min}	Moves the most different phrase to the most different intent

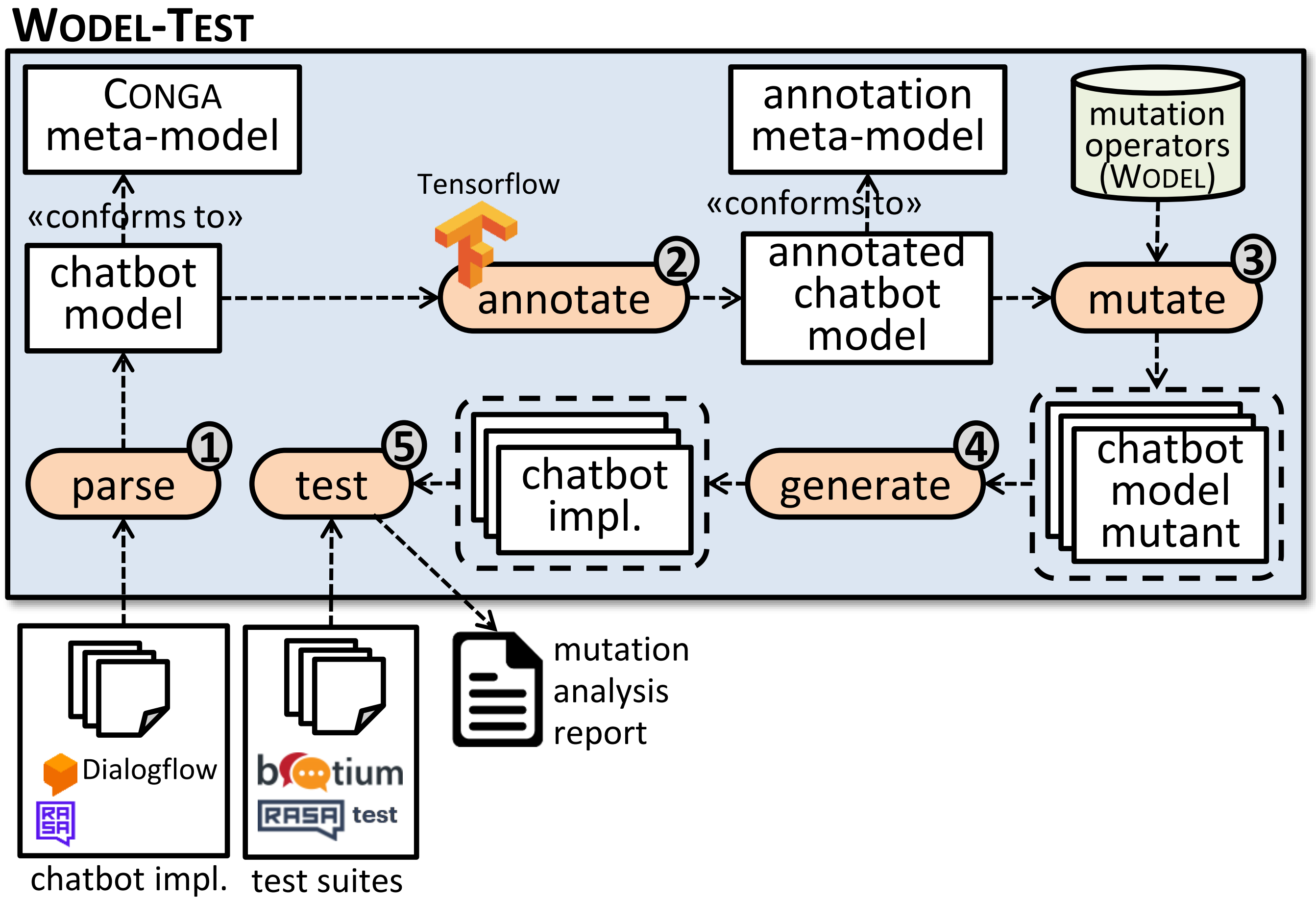
Operators for intents	
DIP	Deletes intent parameter
DPP	Deletes parameter prompt
SPO	Sets required parameter to optional
DFI	Deletes fallback intent
Operators for entities	
CRE	Changes regular expression
DLE	Deletes literal from entity
Operators for actions	
DA	Deletes actions
DPR	Deletes a parameter used in a response
SO	Swaps outputs
Operators for conversation flows	
DCS	Deletes conversation step
DCB	Deletes conversation bifurcation

Emulation of common errors made by chatbot developers

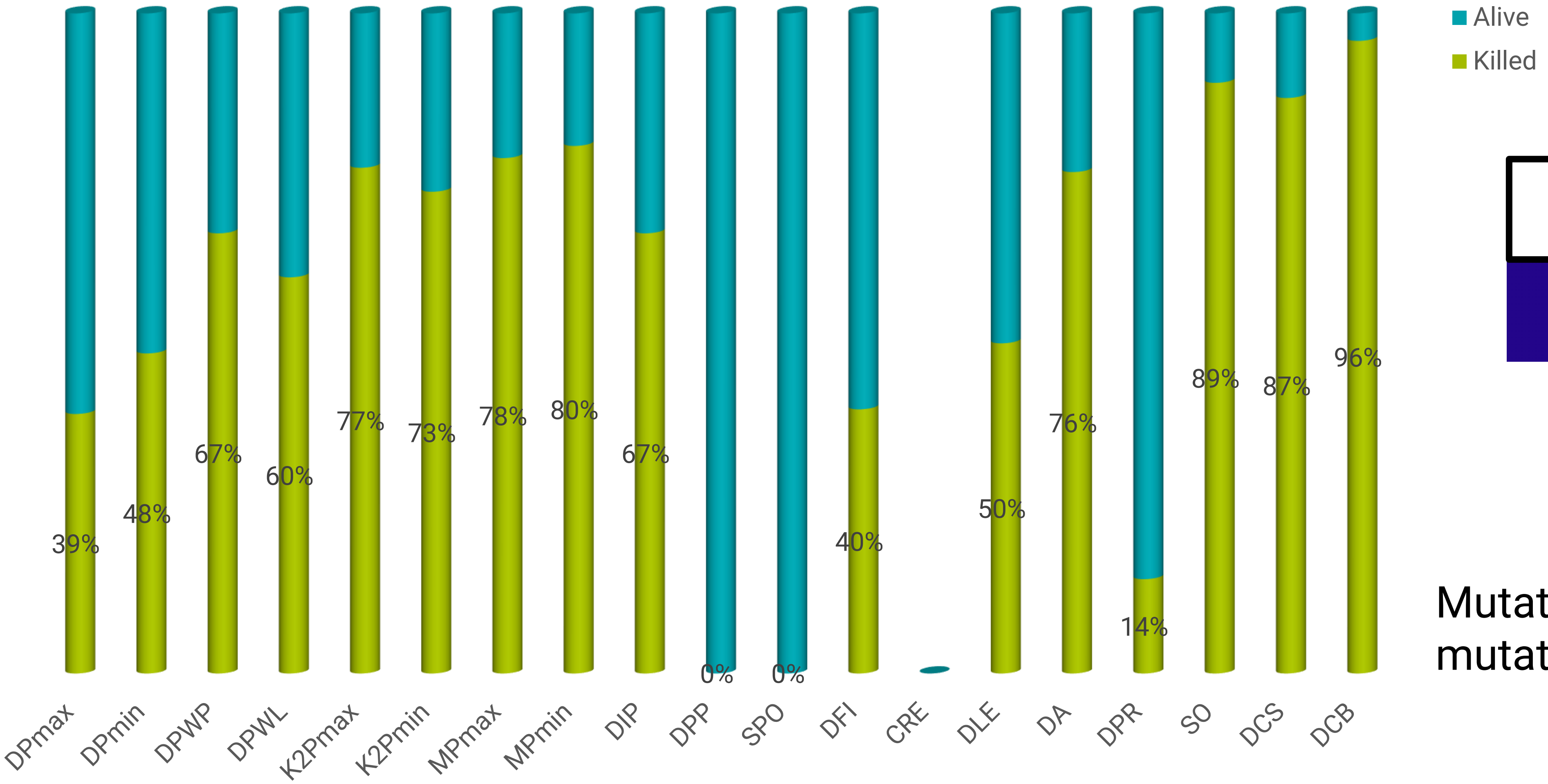
Mutation testing for chatbots

630I

ProxyHands



RQ1: How applicable are the defined mut. ops.?
RQ2: How effective are the defined mut. ops.?

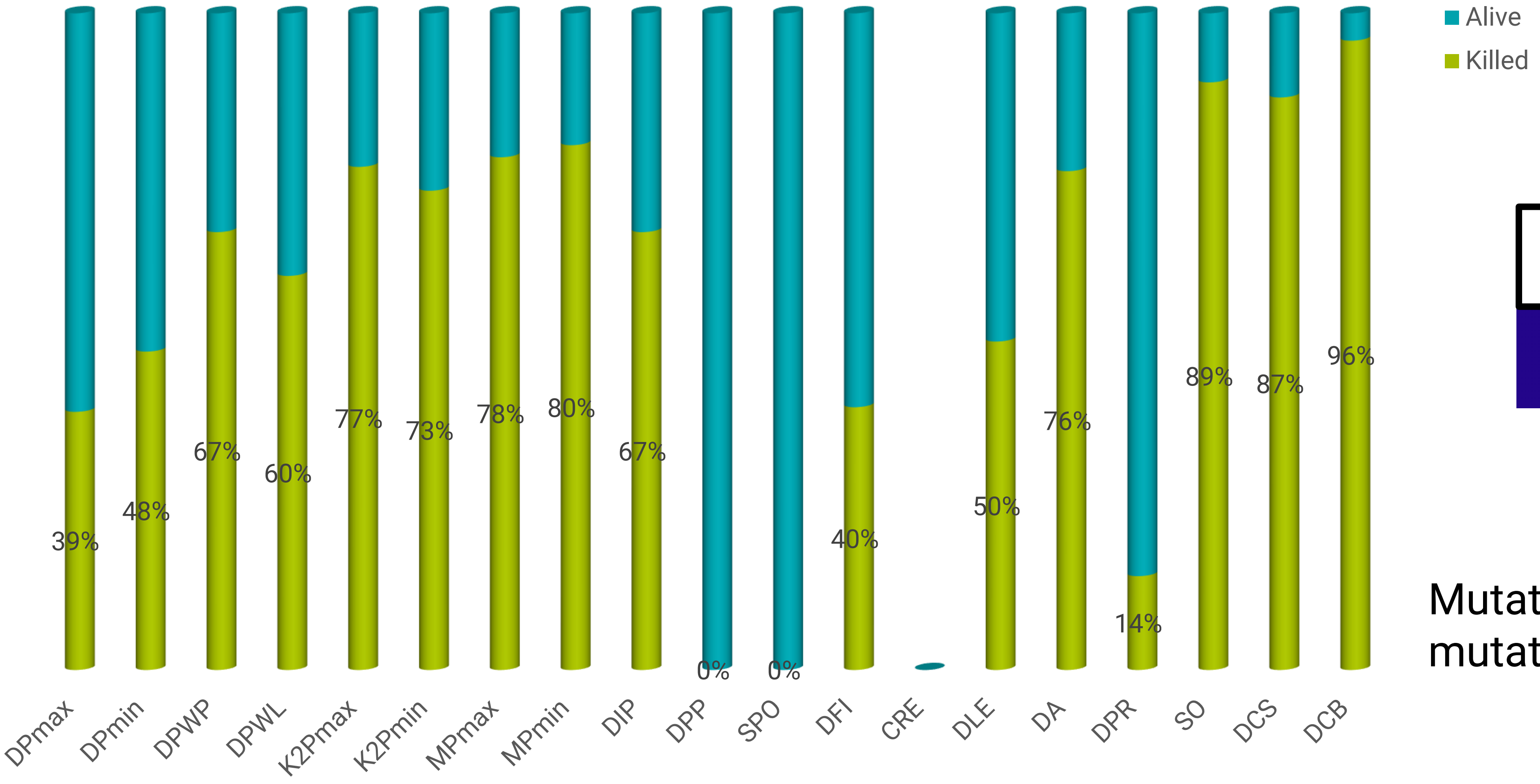


630I

ProxyHands

Mutation score by mutation operator

RQ1: How applicable are the defined mut. ops.?
RQ2: How effective are the defined mut. ops.?

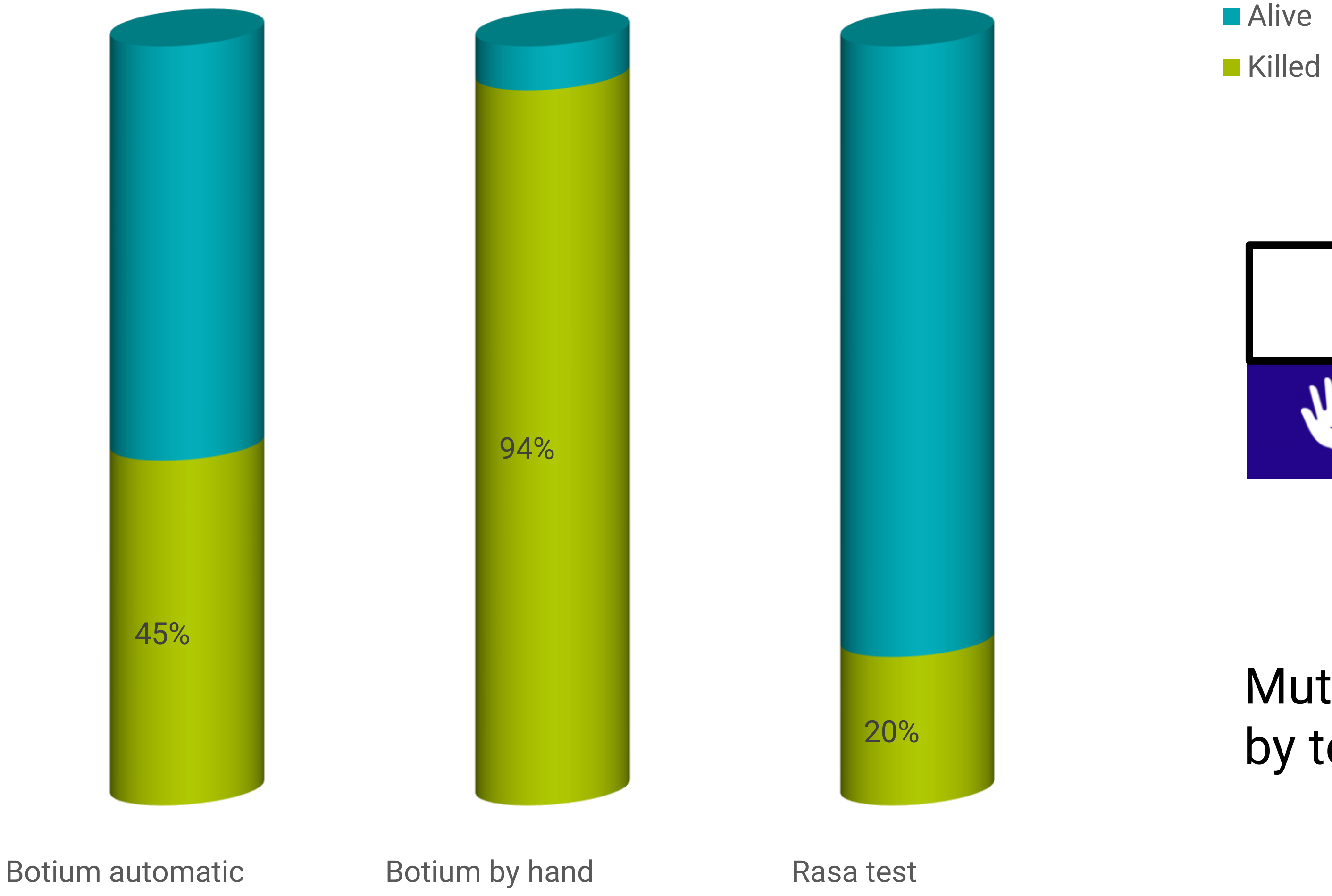


630I

ProxyHands

Mutation score by mutation operator

RQ3: How effective is the MuT process?

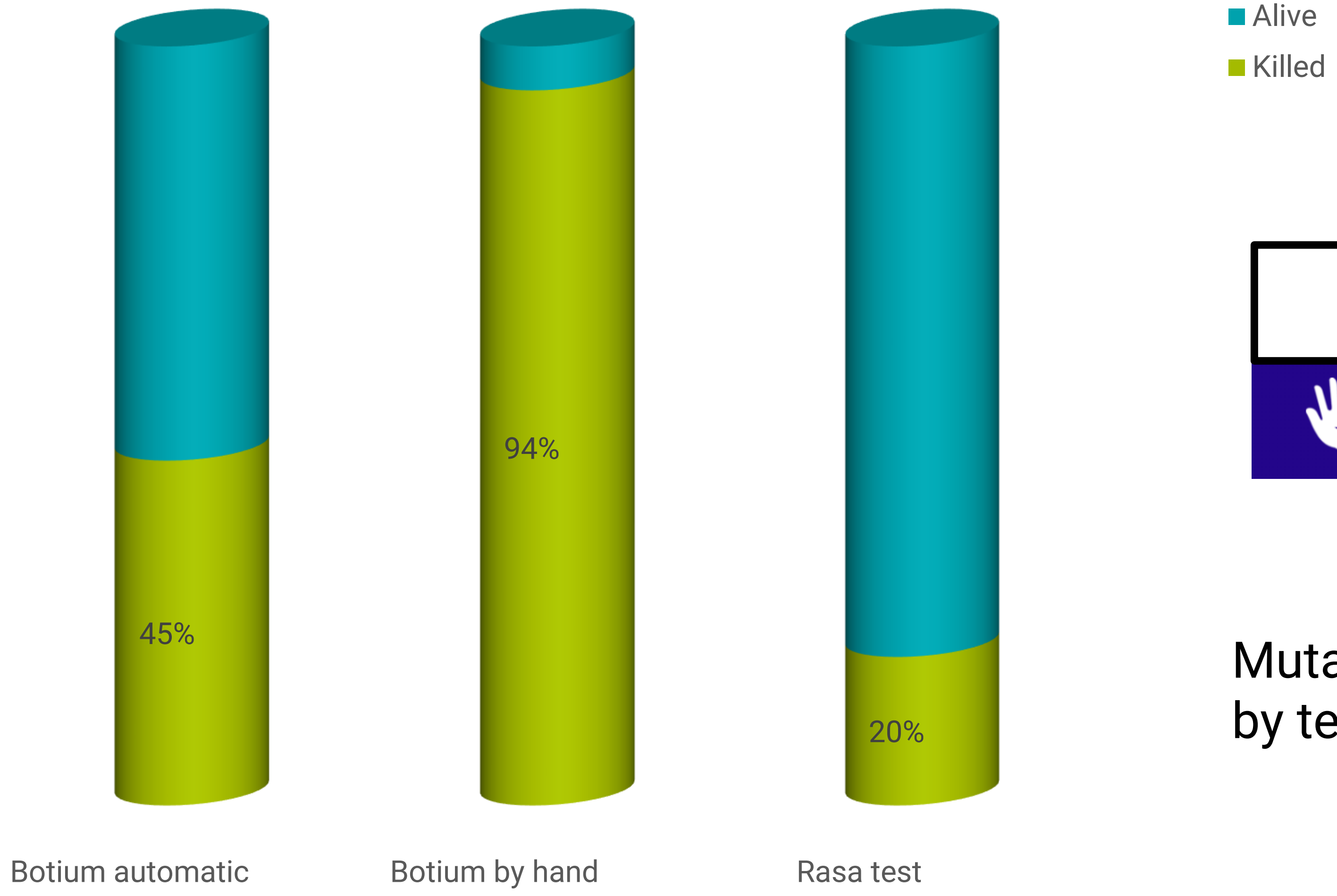


630I

ProxyHands

Mutation score
by test suite kind

RQ3: How effective is the MuT process?

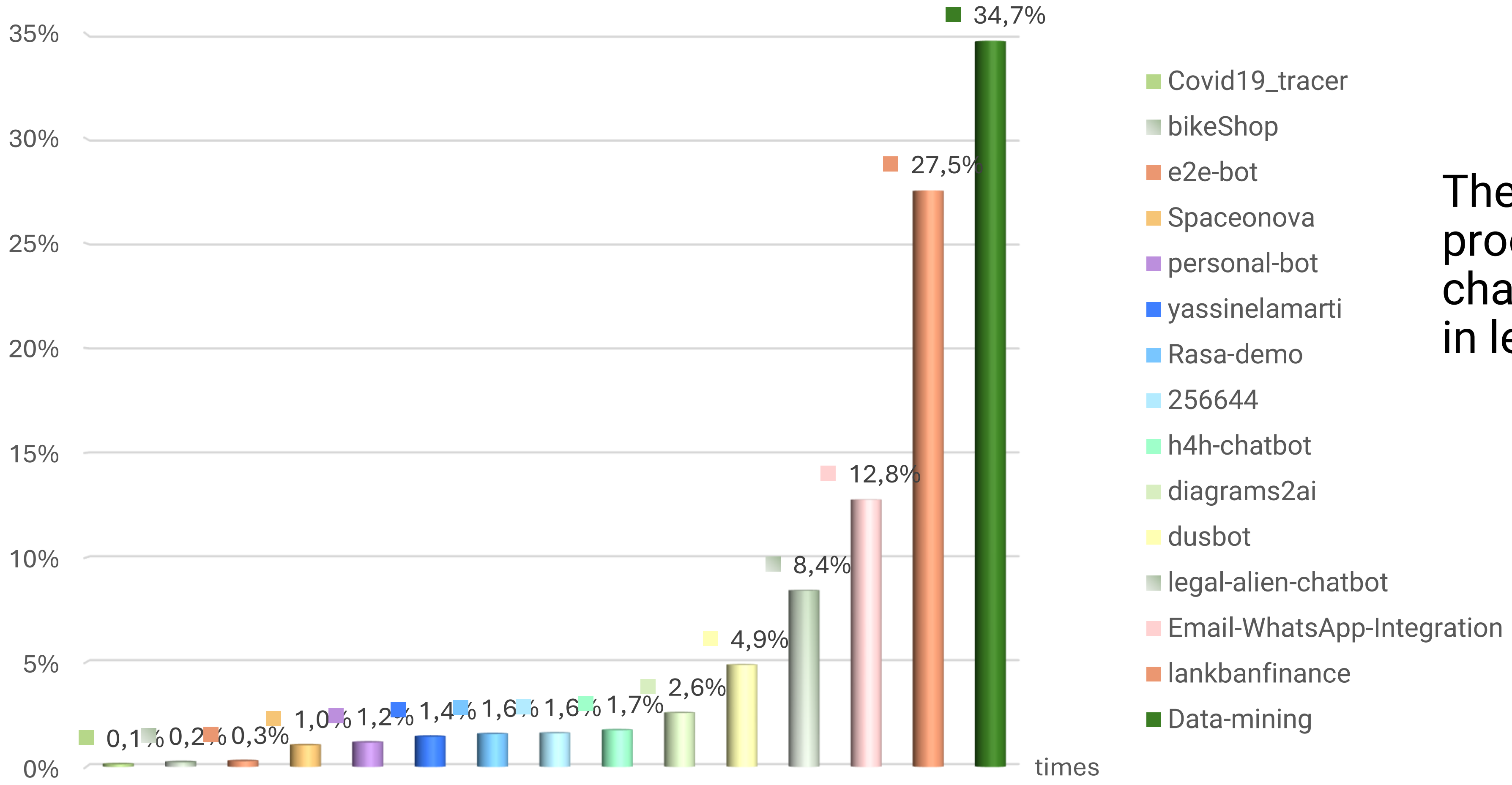


630I

ProxyHands

Mutation score
by test suite kind

RQ4: How efficient is the MuT process?

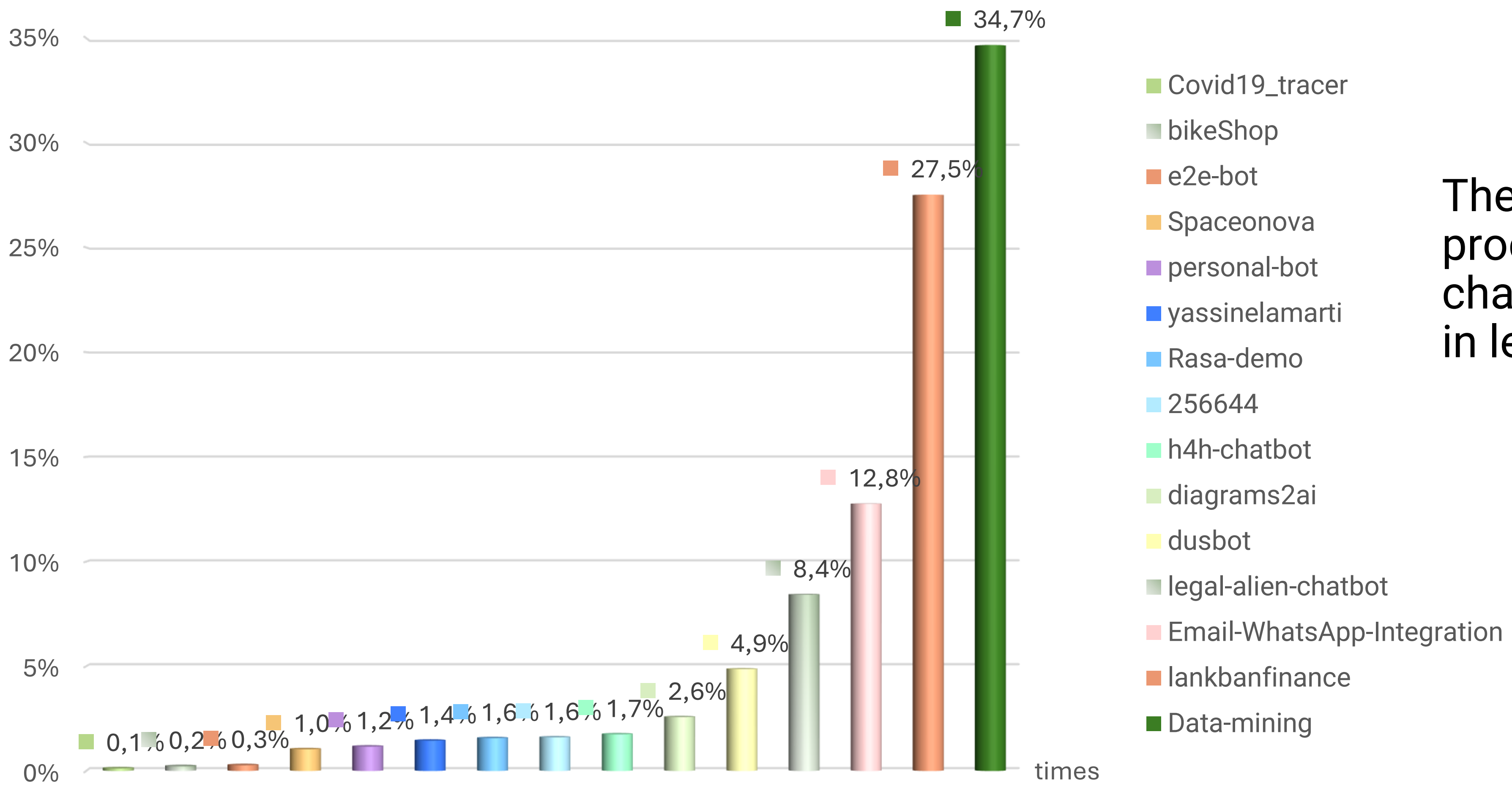


630I

ProxyHands

The mutation testing process of 67% of the chatbots was completed in less than 90 minutes

RQ4: How efficient is the MuT process?



630I

ProxyHands

The mutation testing process of 67% of the chatbots was completed in less than 90 minutes

630I

ProxyHands

Conclusions

- Wodel-Test eases the engineering of MuT tools for DSLs
- Wodel-Test is a better option when we need to
  - Access the source code of the mutants
  - Reason which mutants reduce the mutation score and why
  - Test new mutation operators

Future work

630I

ProxyHands

- Automate the detection of semantically equivalent mutants
  - e.g., in the case of chatbots using confidence decrease heuristics
- Automate the synthesis of tests able to kill the alive mutants
- Optimize the MuT process → Parallelize the mutants generation
- Chatbots: adapt our approach to LLM-based agents

Mutation testing for DSLs

The case of task-oriented chatbots

Pablo Gómez-Abajo

Modelling & Software Engineering Research Group

Universidad Autónoma de Madrid

LangDev
CON 2024

630I

ProxyHands

Seville 17-19 October, 2024

https://langdevcon.org