TUM Data Innovation Lab
X
Steering Lab

Development of an intrinsic motivation complex for an artificial conversational entity

Project Lead: Dr. Ricardo Acevedo Cabra

Scientific Lead: M.Sc. Olena Schüssler

TUM Co-Mentor: Cristina Cipriani

Team: Shreyash Agerwal, Emanuel Deisler, Oğuz Gültepe, Katharina Hermann, Lennart C. Neumann, Nina Schmid, Nicolas Seppich

STEERING LAB
BY HORVÁTH & PARTNERS

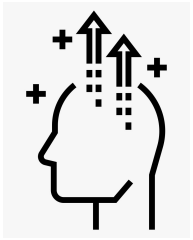
Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

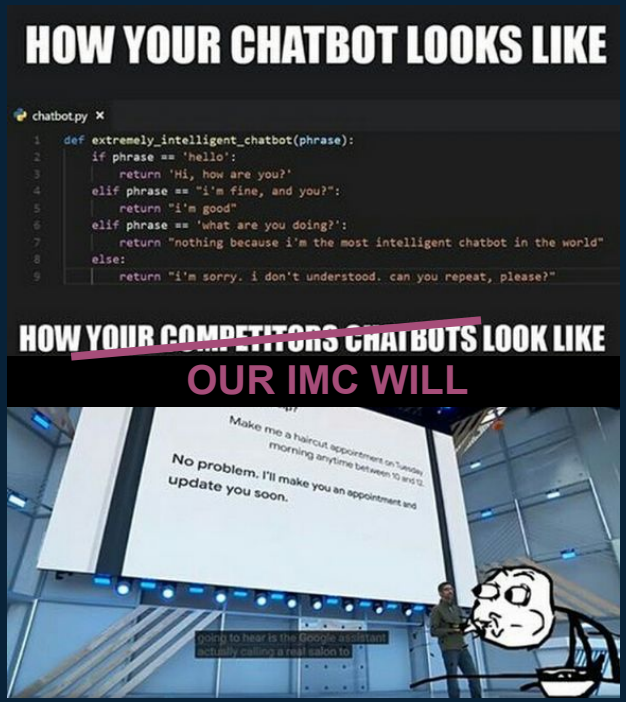
Project Motivation

Today's chatbots ...

- have to be triggered explicitly
- do not adapt autonomously to user needs



Goal:
Intrinsic motivation complex for an artificial conversational entity

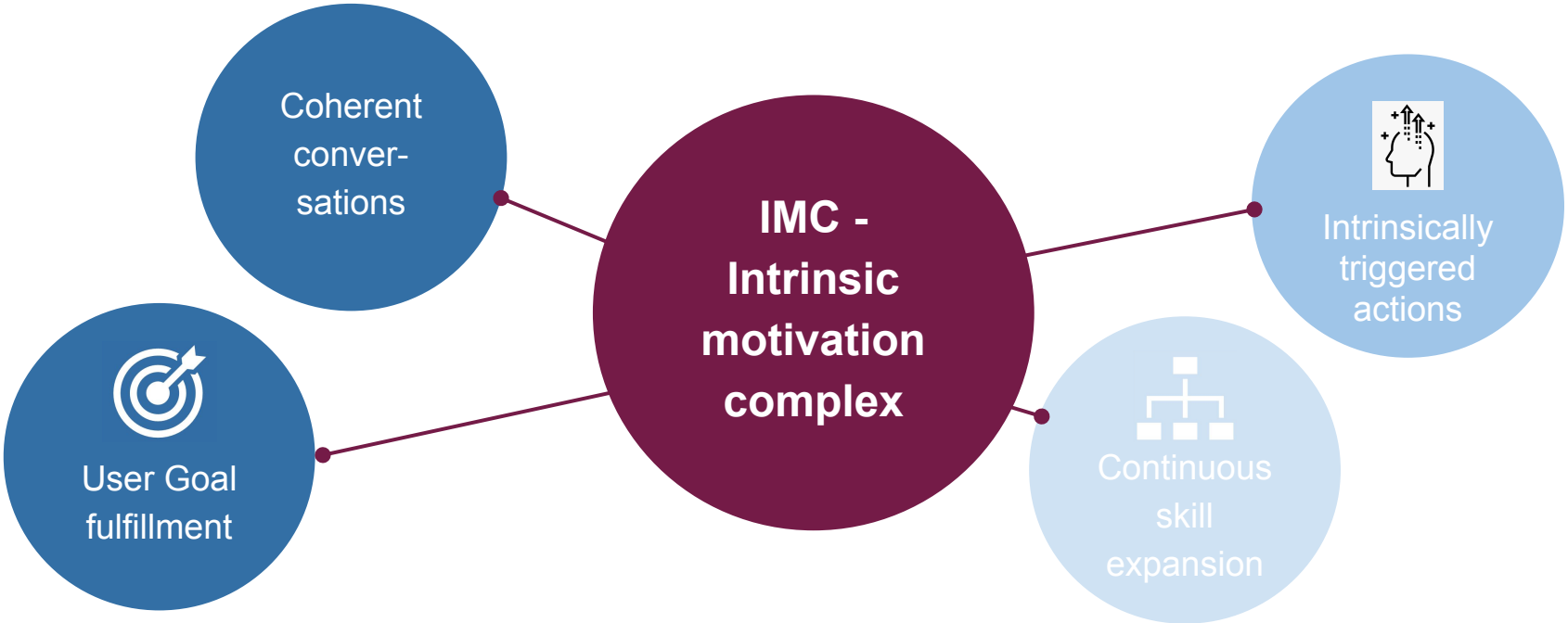


from: https://devrant.com/search?term=chatbots

Project Overview

Overall Objectives

Intrinsic motivation complex for an artificial conversational assistant



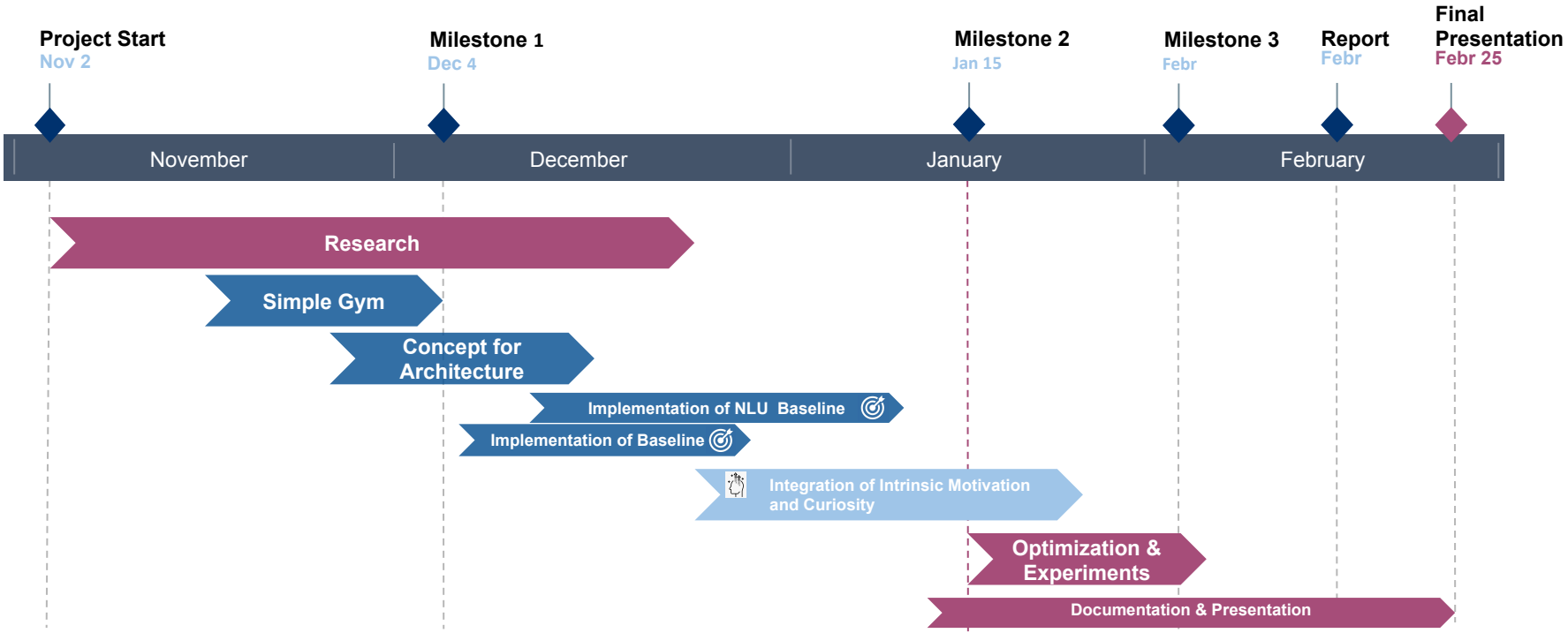
Project Overview

Potential Use Cases

Office Assistant	Bar Tender	Personal buddy
Goal: Schedule day to maximize user's productivity	**Goal:** Maximizing guests' satisfaction by serving the right drinks	**Goal:** Learn user's preferences to maximize its happiness
Agent learns to schedule work packages based on user priorities when being asked	Agent learns to match drinks to user preferences	Agent suggests the user leisure activities; tells him jokes when being asked
Agent learns to suggest the user a new schedule without being asked	Agent learns to suggest drinks on its own without a special order	Agent learns to suggest leisure activities, tell jokes on its own

Project Overview

Project Outline



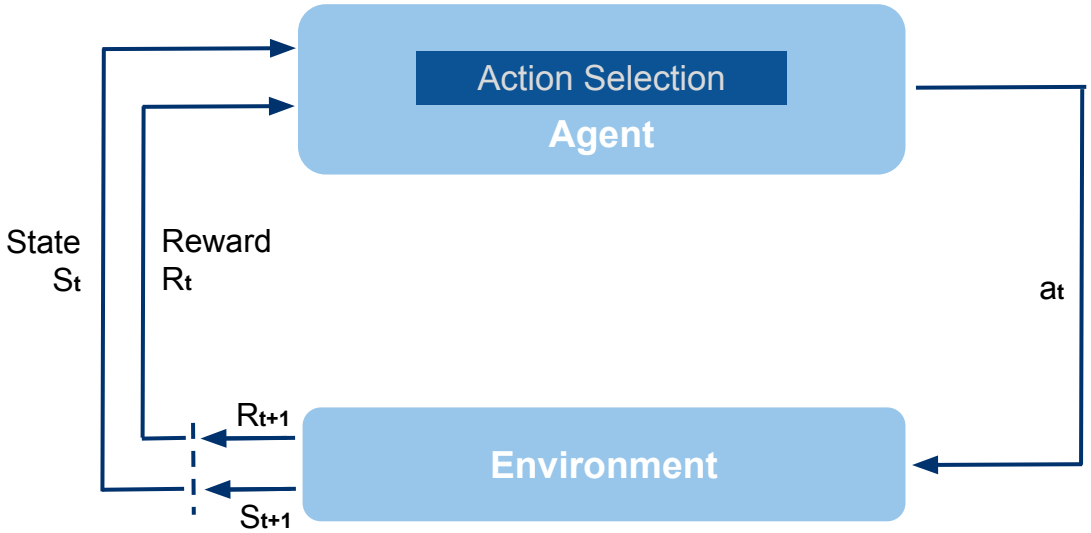
Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

Scientific Concepts

Reinforcement Learning

Reinforcement learning (RL) is an area of machine learning, concerned with how software agents should take actions in an environment to maximize a reward. The agent learns its behavior based on feedback from the environment.



$$a = \pi^*(s)$$

$$\pi^*(s) = \arg\max_{\pi} \sum_t r_t$$

Modern reinforcement learning (Sutton and Barto)

Scientific Concepts

Reinforcement Learning

Classical conditioning (Pavlov, 1960): associates rewards to **events**

Value func:

$$V(s_t) \leftarrow V(s_t) + \alpha(r_t + \gamma V(s_{t+1}) - V(s_t))$$

Instrumental conditioning (Thorndike, 1927; Skinner, 1965): associates rewards to **behaviours.**

Action value func:

$$Q_i(s_t, a_t) \leftarrow Q_i(s_t, a_t) + \alpha[r_t + \gamma max_{a_{t+1}} Q_i(s_{t+1}, a_{t+1}) - Q_i(s_t, a_t)]$$

Bellman equation: $Q^*(s_t, a_t) = E[r_{t+1} + \gamma max_a Q^*(s_{t+1}, a_{t+1})]$

Scientific Concepts

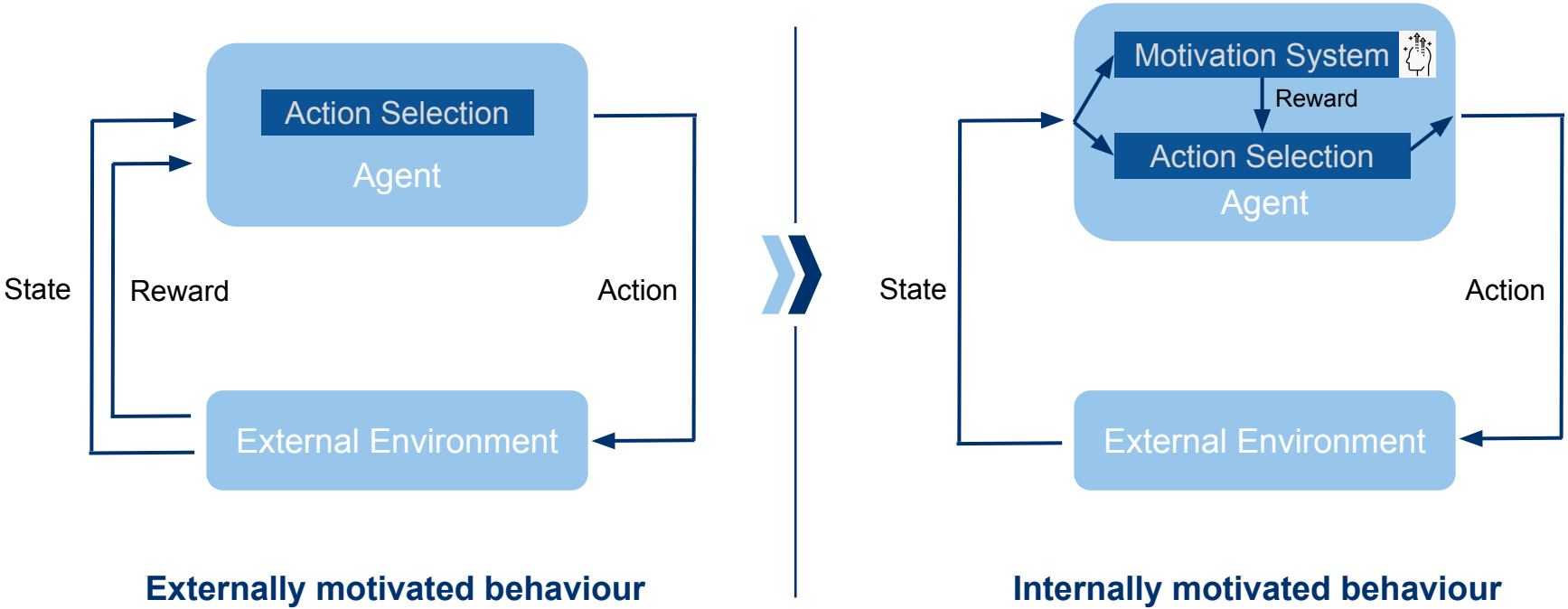
Intrinsic motivation in Psychology

Definition by the American Psychological Association:

*“An incentive to engage in a specific **activity** that **derives** from **pleasure in the activity itself** (e.g., a genuine interest in a subject studied) rather than because of any external benefits that might be obtained (e.g. money, course credits).”*

Scientific Concepts

Intrinsic motivation in Reinforcement Learning



Externally motivated behaviour

Internally motivated behaviour

Scientific Concepts

Intrinsic motivation in Reinforcement Learning

Empowered Agents:
Maximize the mutual information between the **expected outcome of the agent's actions** and the **consequences of its actions** (Gregor et. al. Variational Intrinsic Control (2016)

Curiosity driven learning:
Intrinsic **reward** is equal to the **error** of our agent to **predict the next state** given the current state and **action taken** (Pathak et al Curiosity driven learning 2017)

Scientific Concepts

Advantages and Challenges

Advantages

- Tackling of **sparse rewards** or non-existing rewards problem
- Possibility to **incrementally learn skills** independently of the agent's main task

Challenges for the project

- ✓ Finding a **good policy** and **motivational system**
- ✓ Prioritizing tasks
- ✓ Optimizing for **complex** or **rapid-changing observations**
- ✓ Implementing a **good user simulation**

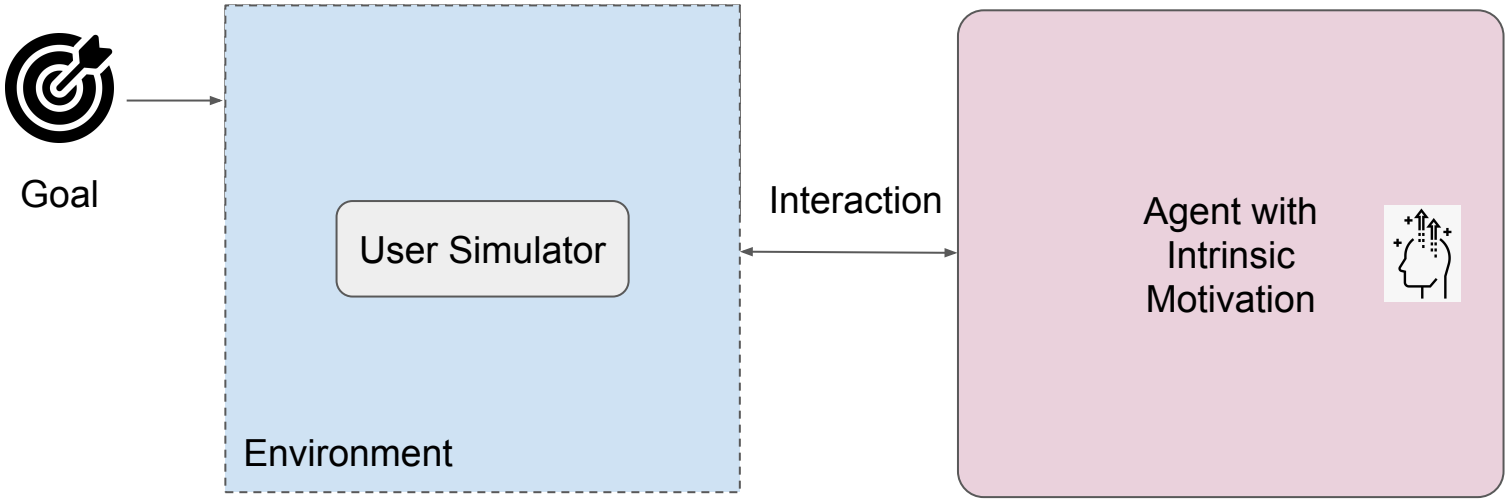
Andrew et al . Policy invariance under reward transformations: Theory and application to reward shaping. (1999)
Aubret et al. A survey on intrinsic motivation in reinforcement learning. (2019)

Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

Baseline

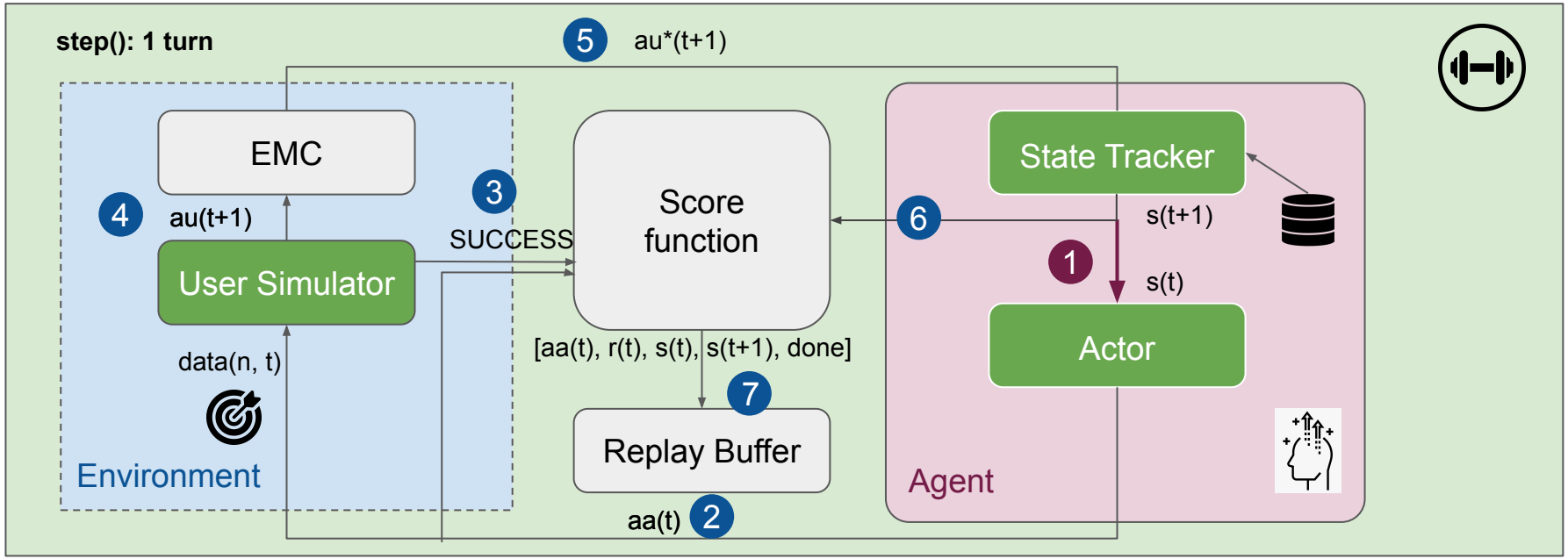
Intrinsic motivated RL agent interacting with a user simulator having specific goals in a discrete world of intents



Baseline

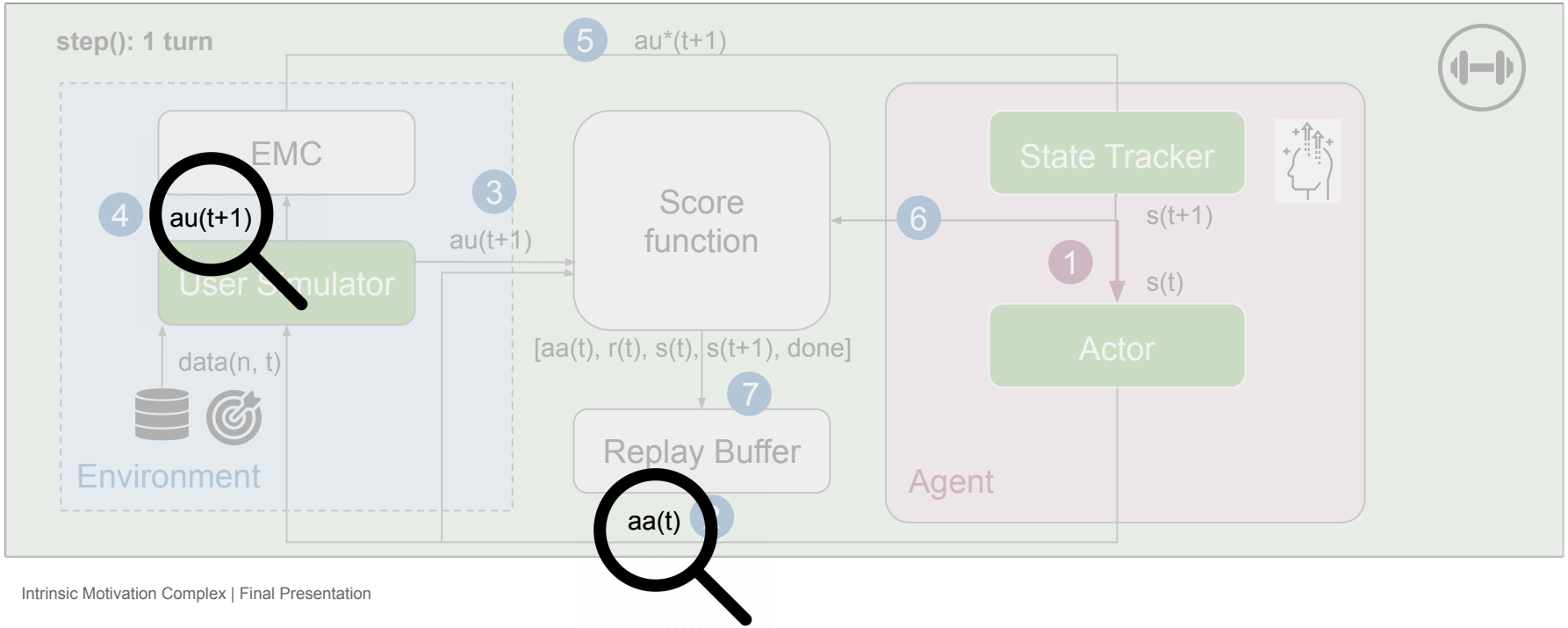
Baseline

Experience Collection Step



Baseline

User & Agent Actions



Baseline

User & Agent Actions

Action Form: {‘intent’: ‘INTENT’, ‘inform_slots’: ‘{‘Slot1’:’Value1’, ..}’, ‘request_slots’:{‘Slot1’: ‘UNK’, ..}

= purpose of an action

user_intents:
order_drinks, inform, request, reject, thanks, goodbye

agent_intents:
utter_request, utter_inform, find_drink, utter_goodbye

= variables from domain

inform slots:
Values to be informed e.g. size of an drink

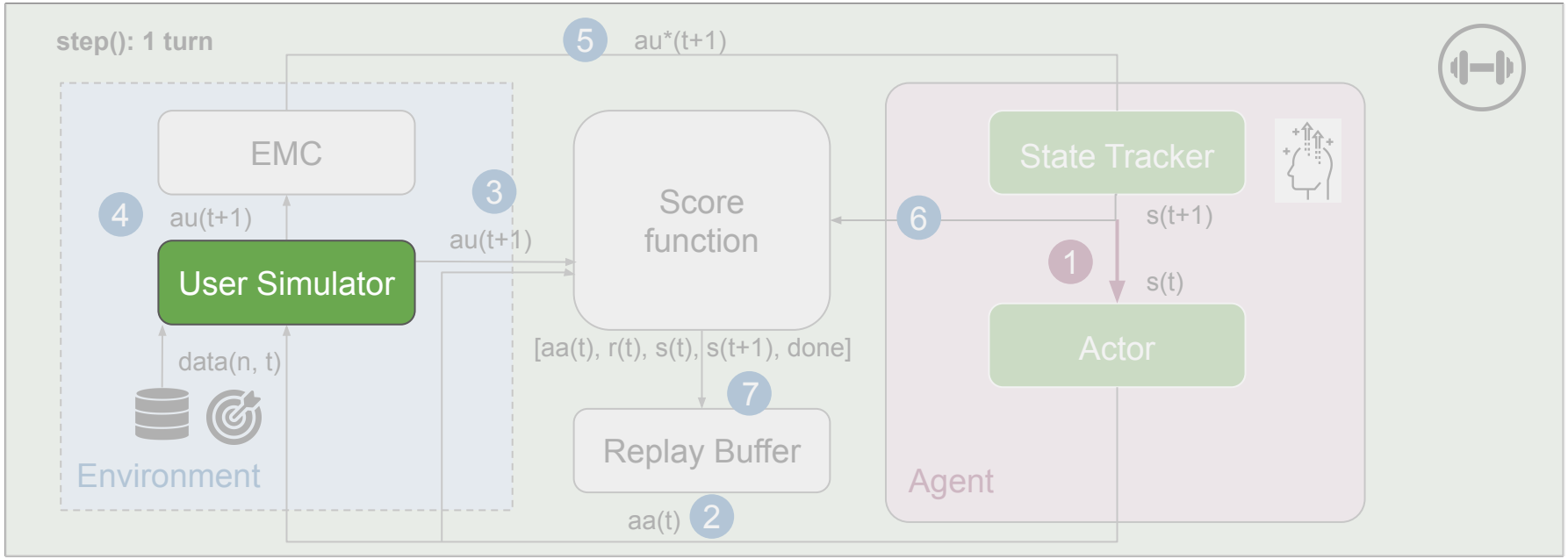
request slots:
Additional values requested, e.g. which size is available

Slot Domain: {DRINK, TEMP, SIZE}

Baseline

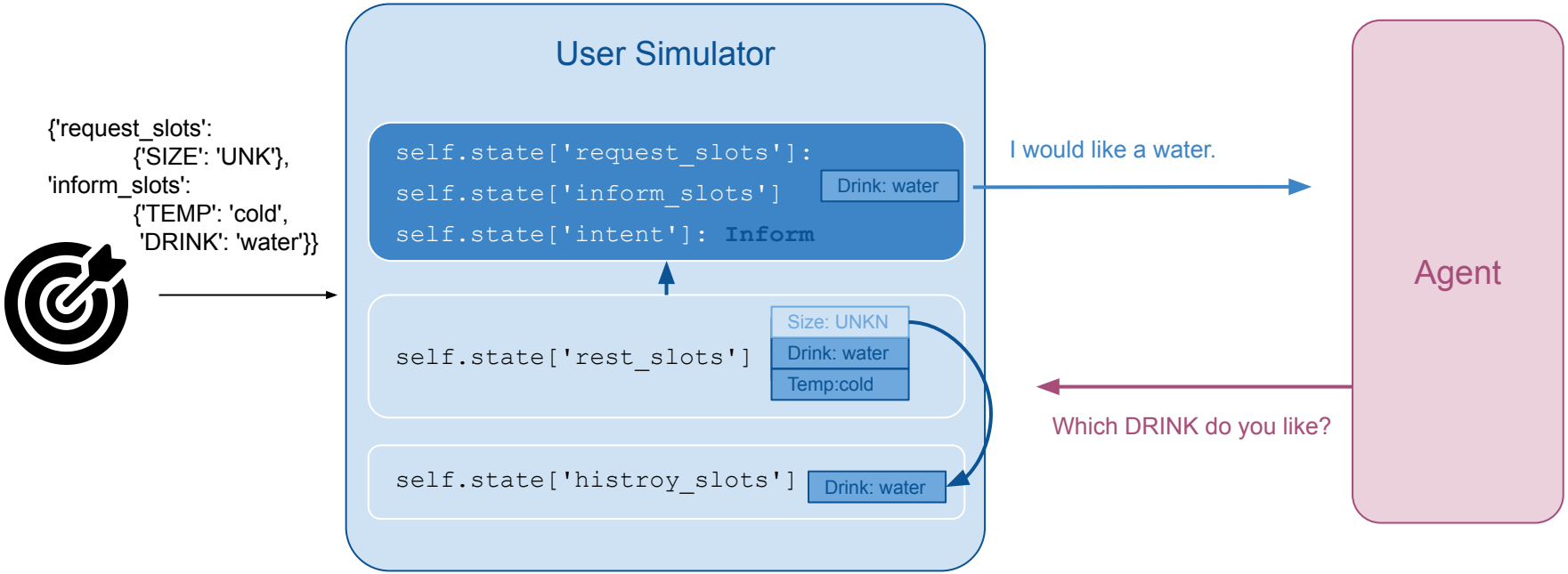
Baseline

Gym: Place, where the agent can explore its state action space and learns, implemented in 3 functions



Baseline

User Simulator



Results for Baseline

Baseline Example Conversation

```
***Episode 23 *******
User Goal:{'request_slots': {'drinknumber': 'UNK'}, 'diaact':'request', 'inform_slots':{'DRINK': 'cola', 'SIZE': 'small'}}
--------------------------------------------------------------------------------
Initial User Utterance: {'intent': 'order_drinks', 'request_slots': {}, 'inform_slots': {}, 'round': 0, 'speaker': 'User'}
Agent Action: {'intent': 'utter_request', 'inform_slots': {}, 'request_slots': {'DRINK': 'UNK'}, 'round': 1, 'speaker': 'Agent'}

User Response: {'intent': 'inform', 'request_slots': {}, 'inform_slots': {'DRINK': 'cola'}}
Agent Action: {'intent': 'utter_inform', 'inform_slots': {'SIZE': 'small'}, 'request_slots': {}, 'round': 2, 'speaker': 'Agent'}
User Response: {'intent': 'request', 'request_slots': {'drinknumber': 'UNK'}, 'inform_slots': {}}
Agent Action: {'intent': 'find_drink', 'inform_slots': {'DRINK': 'cola', 'SIZE': 'small', 'drinknumber': '0'}, 'request_slots': {}, 'round': 3, 'speaker': 'Agent'}
User Response: {'intent': 'thanks', 'request_slots': {}, 'inform_slots': {}}
Agent Action: {'intent': 'utter_goodbye', 'inform_slots': {}, 'request_slots': {}, 'round': 4, 'speaker': 'Agent'}
User Response: {'intent': 'goodbye', 'request_slots': {}, 'inform_slots': {}}
Episode: 23 Success: True Reward: 6
```

Example:

- 2 slots scenario: DRINK and SIZE
- No Rendering

Result:

- Learns simple conversation
- Works with 98% success rate

Agenda

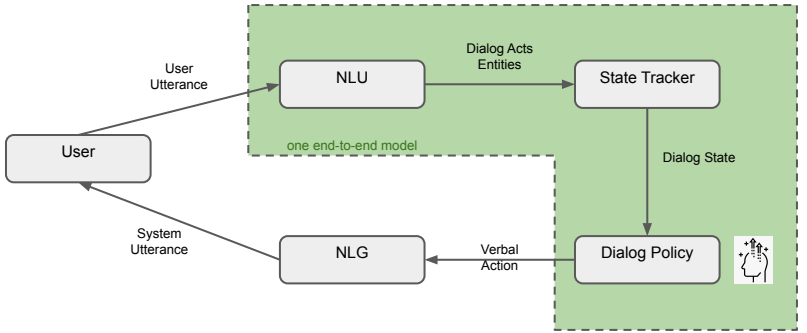
1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

End-to-end pipeline of a task-oriented and intrinsic spoken dialog system

NLU approaches

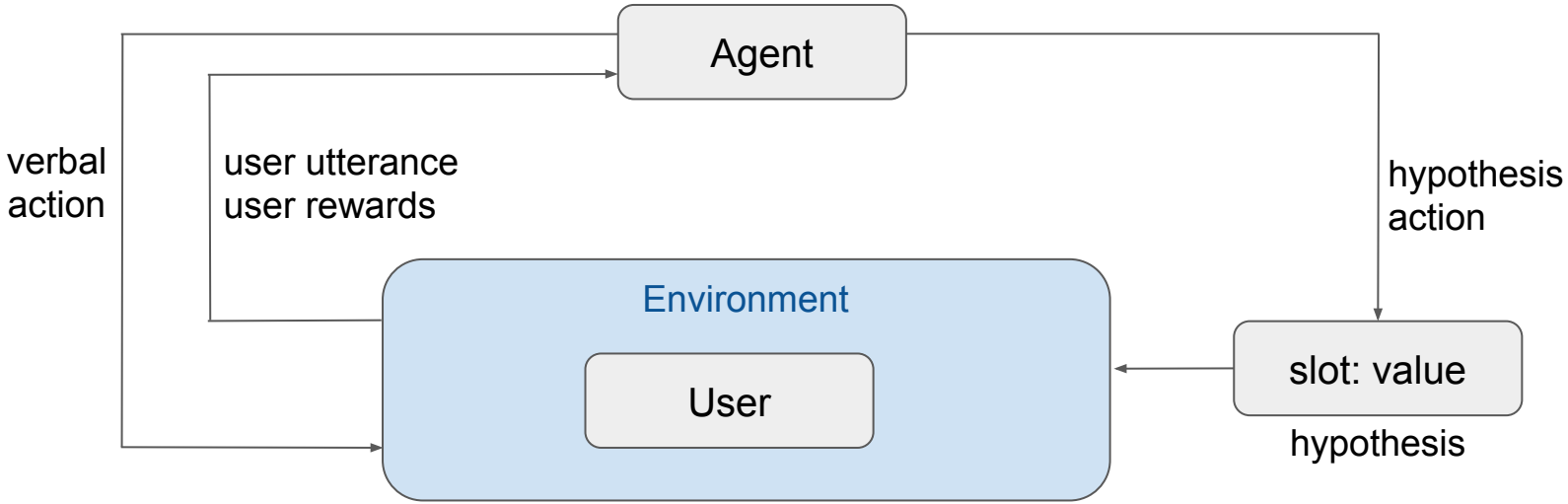
a multicomponent approach

b end-to-end approach



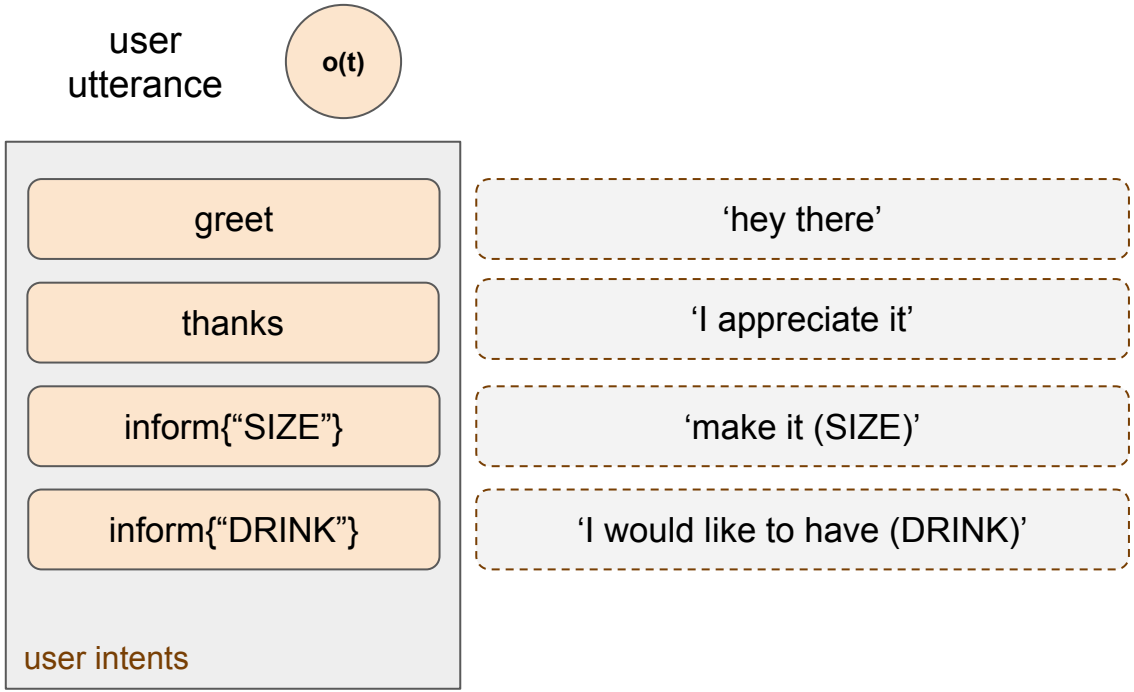
NLU: Multicomponent Approach

End-to-end pipeline of a task-oriented spoken dialog system



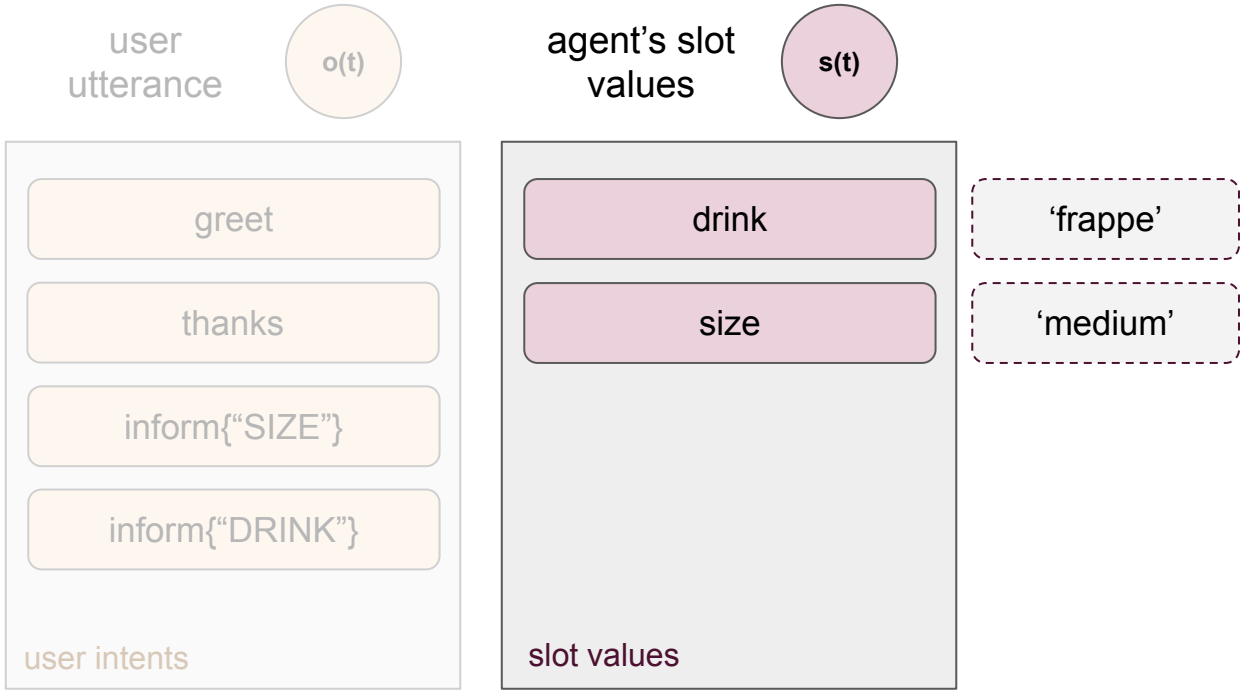
NLU: Multicomponent Approach

Pipeline Components



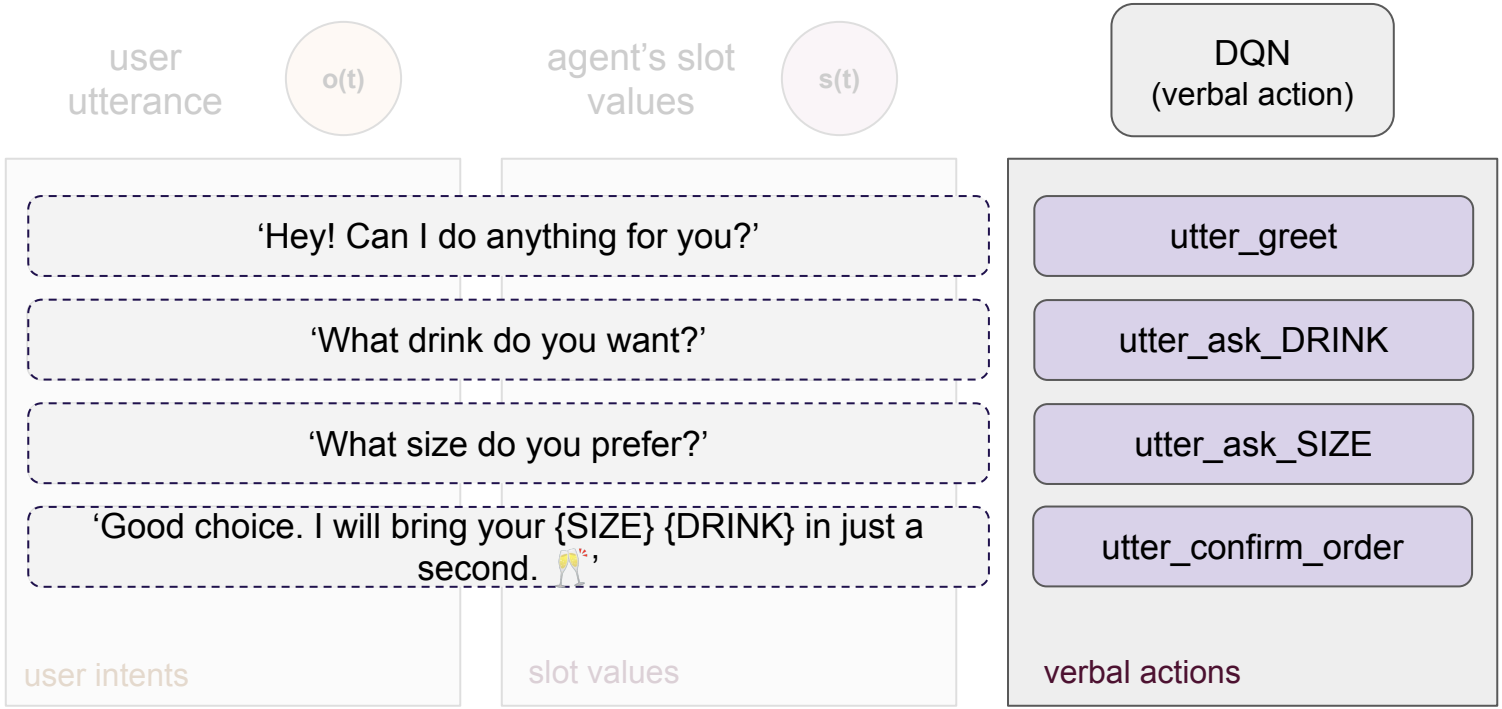
NLU: Multicomponent Approach

Pipeline Components



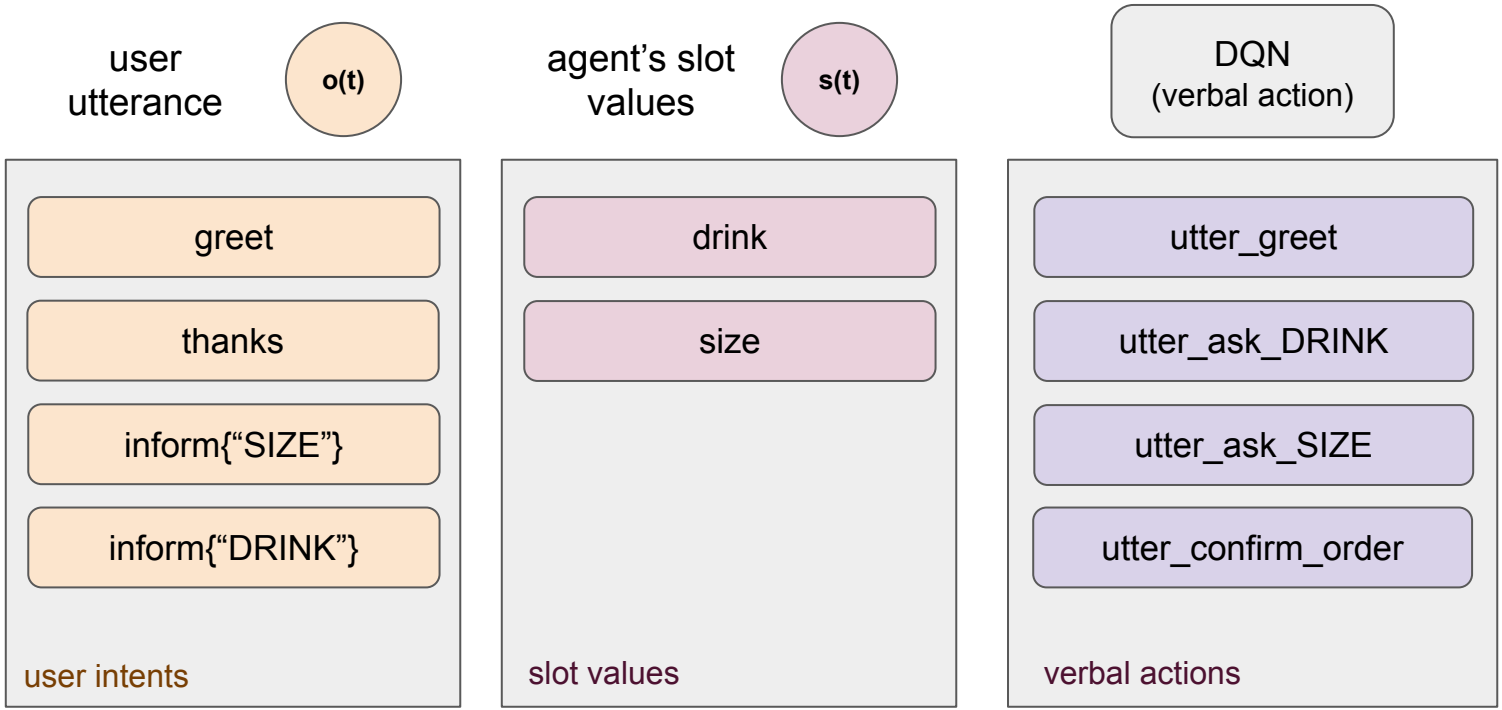
NLU: Multicomponent Approach

Pipeline Components



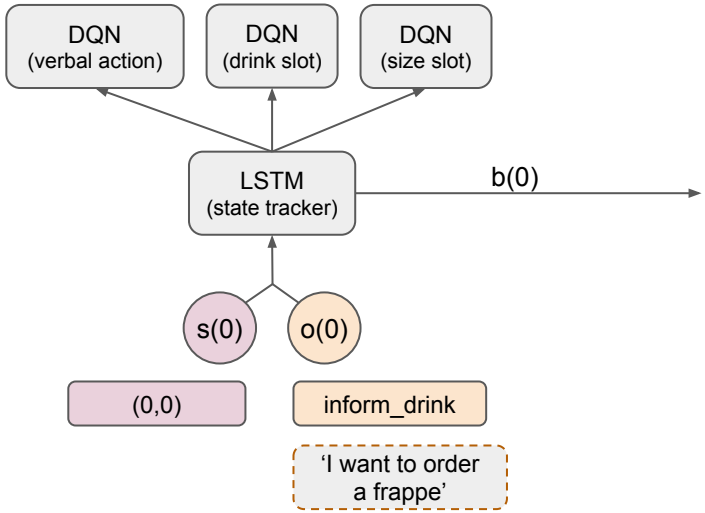
NLU: Multicomponent Approach

Pipeline Components



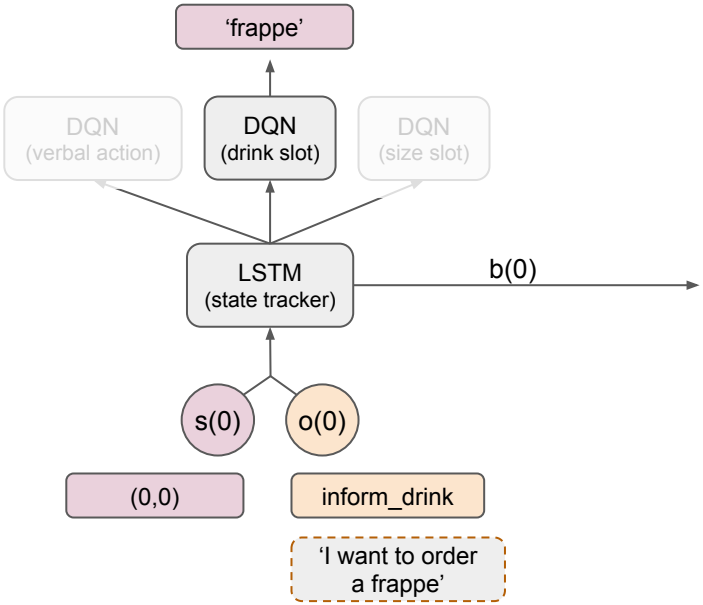
NLU: Multicomponent Approach

Example of an Episode



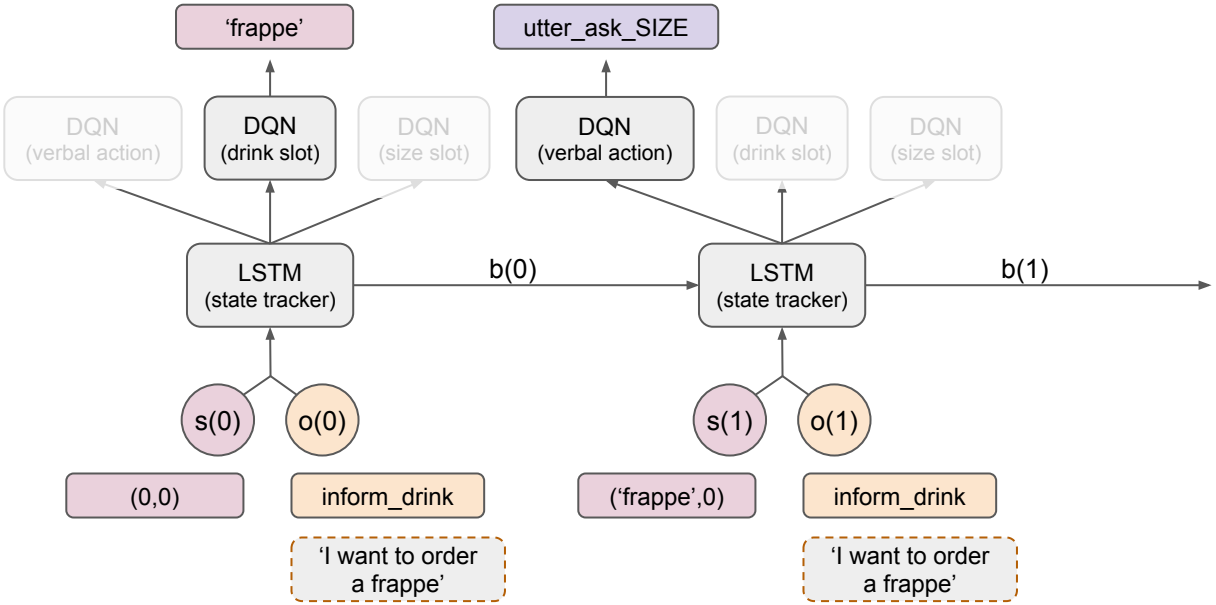
NLU: Multicomponent Approach

Example of an Episode



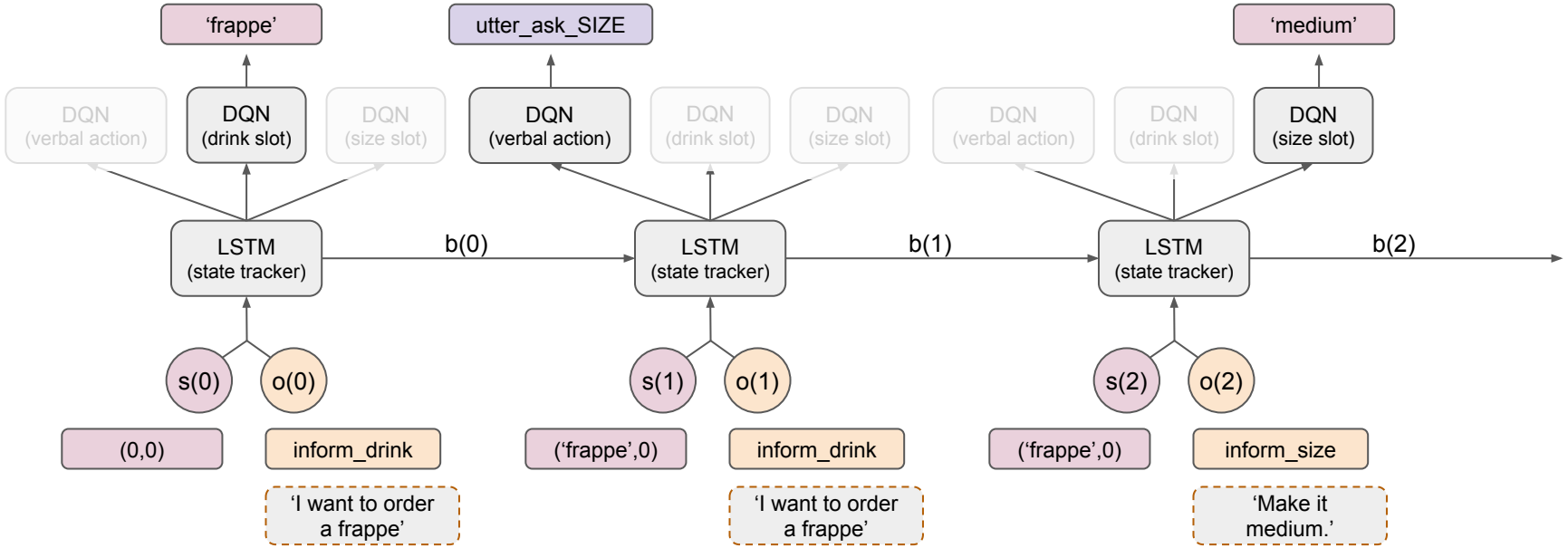
NLU: Multicomponent Approach

Example of an Episode



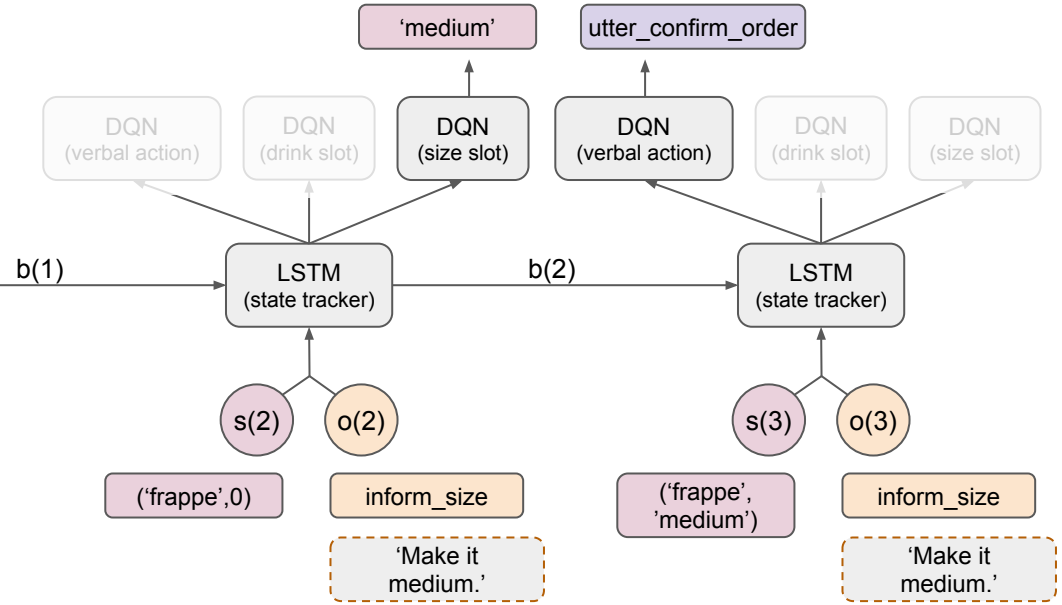
NLU: Multicomponent Approach

Example of an Episode



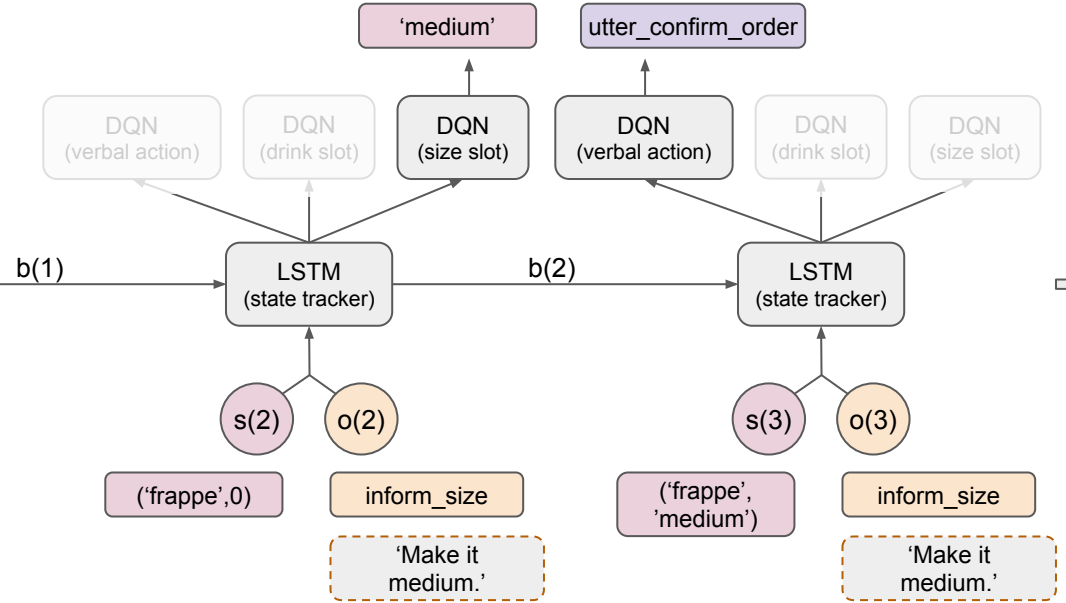
NLU: Multicomponent Approach

Example of an Episode



NLU: Multicomponent Approach

Example of an Episode



Termination of the episode

if agent's slot values equal slot values of user goal → positive reward

else → negative reward

NLU: Multicomponent Approach

Results

Best result: 15% episode success rate
(for a simplified database containing a single story, 20 drinks and 10 verbal utterances)

Longer training, small implementation changes, etc. did not improve that result any further.

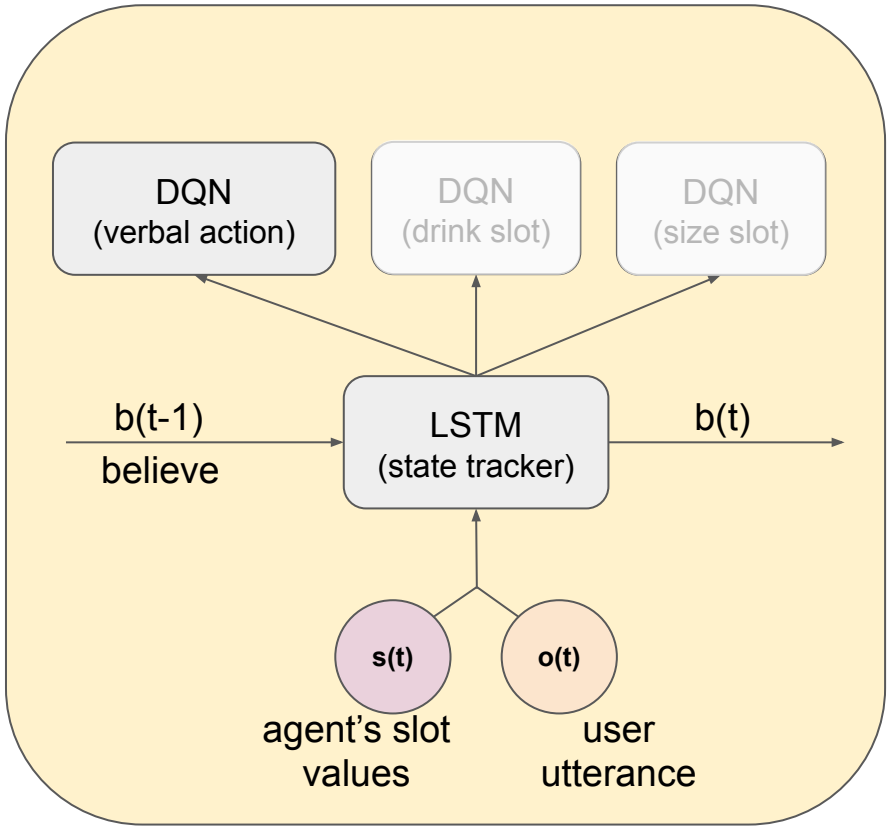
```
********** Episode 1676 ********************
User Goal: 191 , 193
------------------------------------------------------------
Initial User Utterance: i would like to have fruit cooler
Agent Action: fruit cooler
Agent Action: What size do you prefer?
User Response: oh actually make it medium
Agent Action: large

********** Episode 1677 ********************
User Goal: 186 , 194
------------------------------------------------------------
Initial User Utterance: i would like to have frapp
Agent Action: frapp
Agent Action: What size do you prefer?
User Response: oh actually make it large
Agent Action: large
Agent Action: What can I get you? :)
User Response:
```

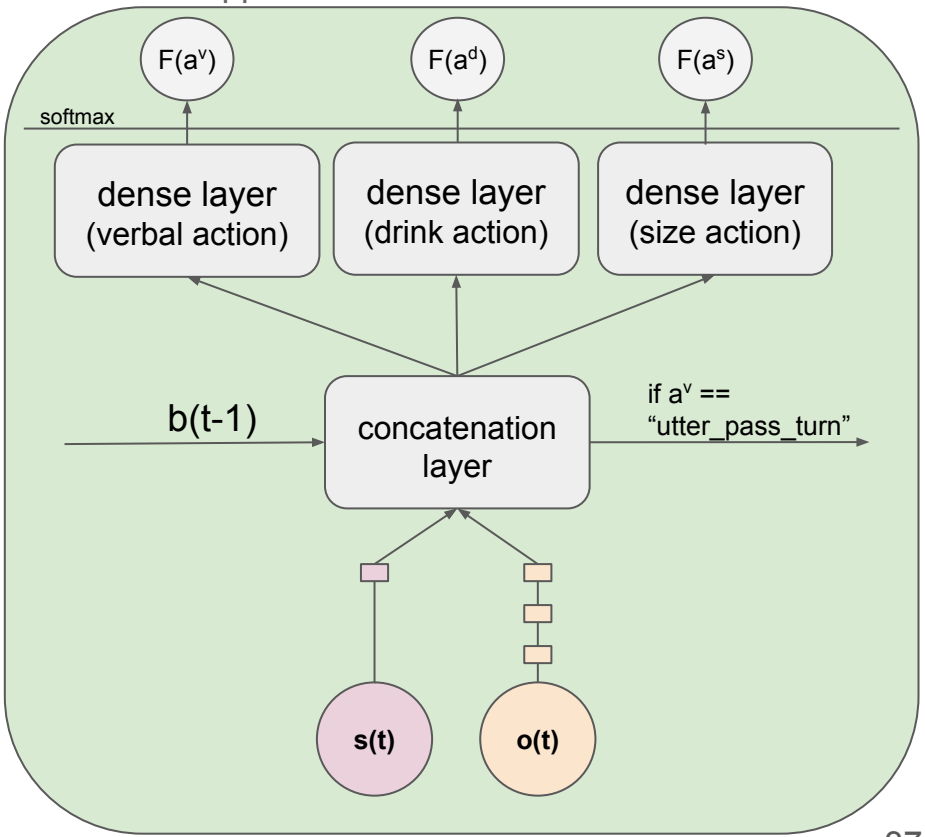
Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

multicomponent approach



end-to-end approach



NLU: End-to-End Approach

Policy networks with 2 layers:

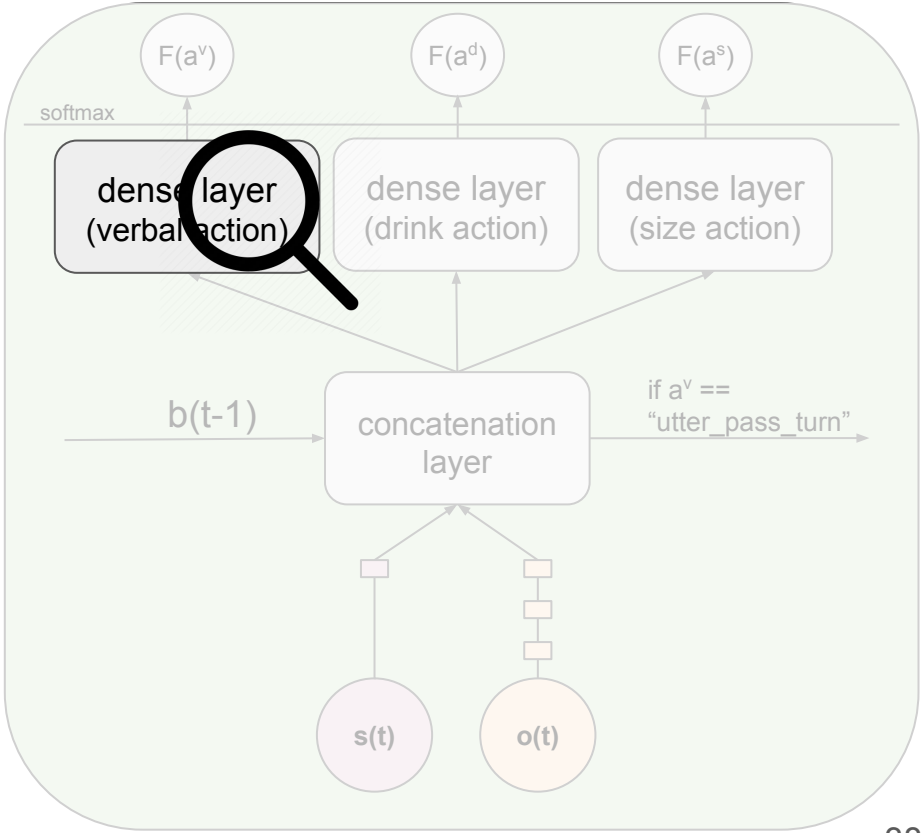
initial layer: size 200

2nd layer: size corresponding to action space

number of possible actions

a^v = 5

- utter_greet
- utter_ask_DRINK
- utter_ask_SIZE
- utter_confirm_order
- **utter_pass_turn**



NLU: End-to-End Approach

Supervised Training Results

training on	lrz.xlarge instance 10 cores batchsize: 4096 Adam optimiser
time needed	~ 5 hours

Categorial Cross Entropy: 1.011
Categorial Accuracy: 0.753

Precision: 1.0
Recall: 0.751

```
#
* greet
    - utter_greet
    - pass_turn
* order_drinks
    - utter_ask_DRINK
    - pass_turn
* inform{"DRINK"}
    - slot{"DRINK"}
    - utter_ask_SIZE
    - pass_turn
* inform{"SIZE"}
    - slot{"SIZE"}
    - utter_confirm_order
```

a sample story

Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

Comparison of the concepts

criteria	intent based approach	NLU approaches	
		multicomponent approach	end-to-end approach
allows for NLU	not yet	yes	yes
performance of the base model	very good (success rate of 98%)	not sufficient	good (acc. of 75%)
possibility of including IM	yes (already included)	probably	probably
easy extendability	yes	yes	moderate
necessary training resources	normal CPU -> 1-2h of training	normal CPU -> 1-2h of training	LRZ cloud -> 5h of training

Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

Baseline Extension

Extended User & Agent Actions

Action Form: {‘intent’: ‘INTENT’, ‘inform_slots’: ‘{‘Slot1’:’Value1’, ..}’, ‘request_slots’:{‘Slot1’: ‘UNK’, ..}

= purpose of an action

user_intents:
order_drinks, inform, request, reject, thanks, goodbye, 'nothing', 'not_helpful'

agent_intents:
utter_request, utter_inform, find_drink, utter_goodbye, trigger_user, joke, utter_nothing'

= variables from domain

inform slots:
Values to be informed e.g. size of an drink

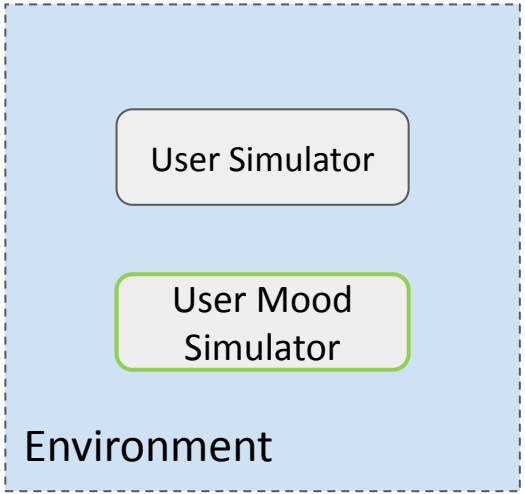
request slots:
Additional values requested, e.g. which size is available

Slot Domain: {DRINK, TEMP, SIZE}

Concept Mood Based IM

Introduction of the User‘s Mood

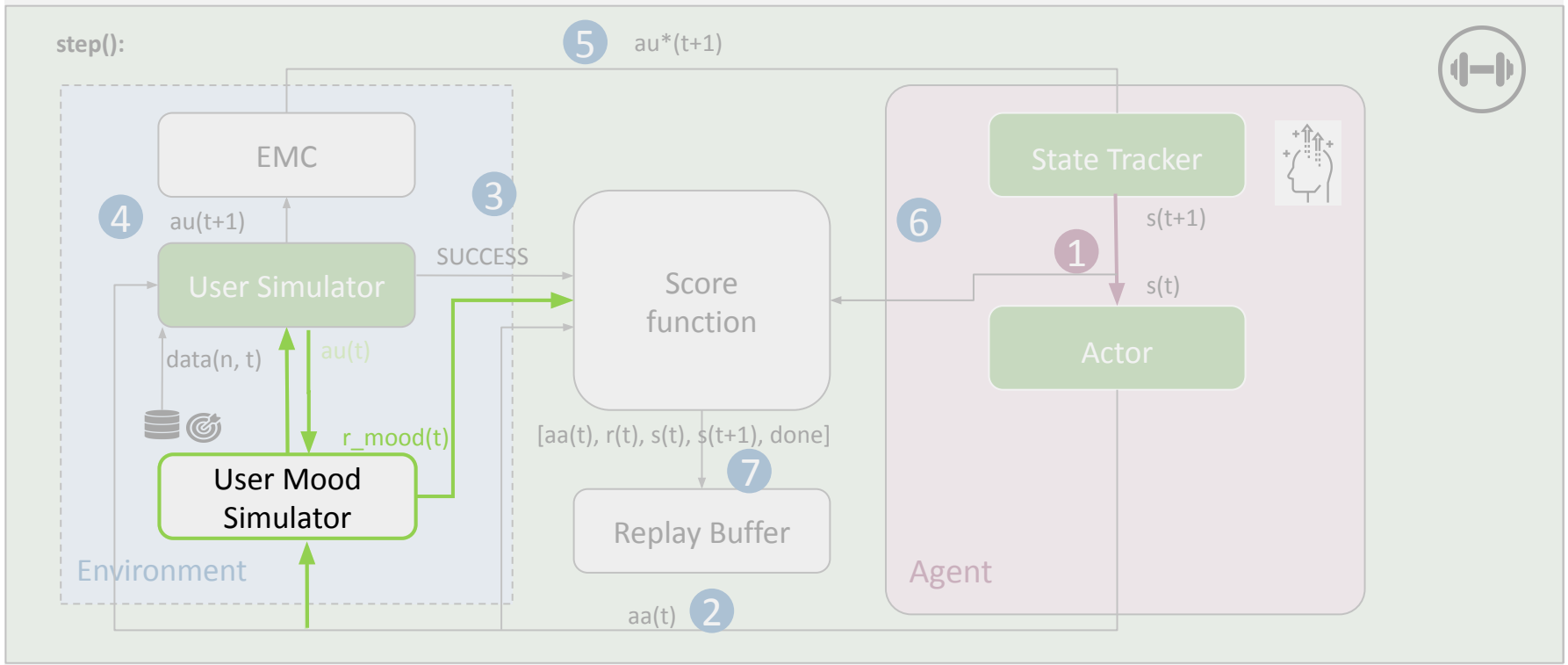
Goal



Agent with Intrinsic Motivation

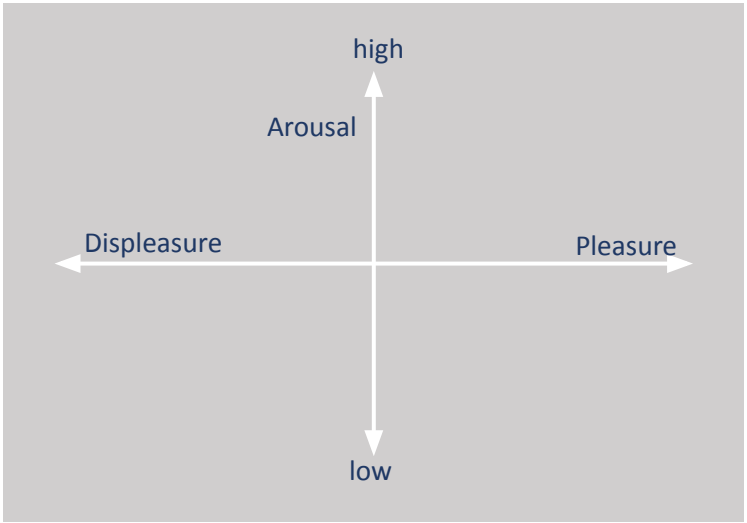
Concept Mood Based IM

How to train the agent



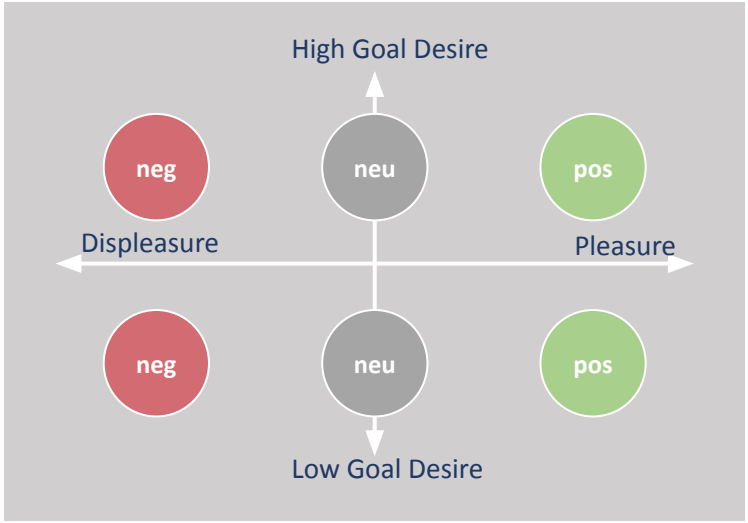
Concept Mood Based IM

User Mood Model



Simplified Two dimensional theoretical emotions model of Russell (PAD emotional state model)*

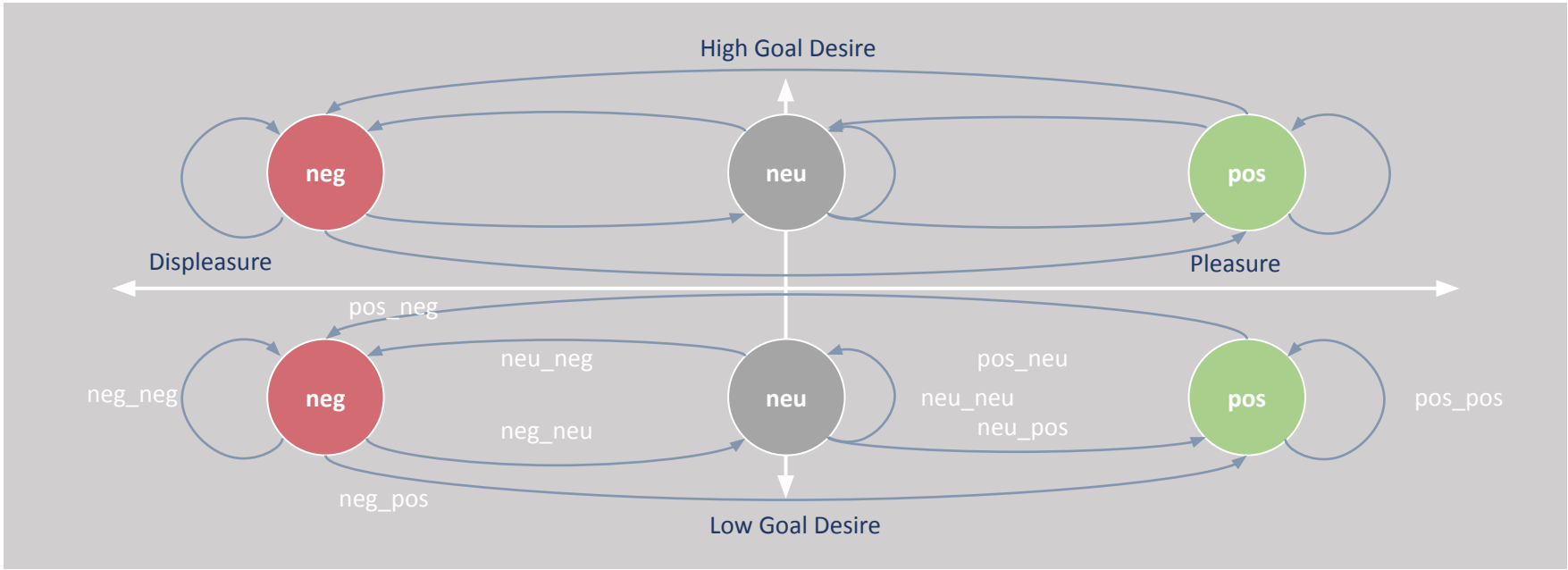
* PAD emotional state model - Wikipedia



Our two dimensional Mood Graph
Which can only have transitions on one level of goal desire currently.

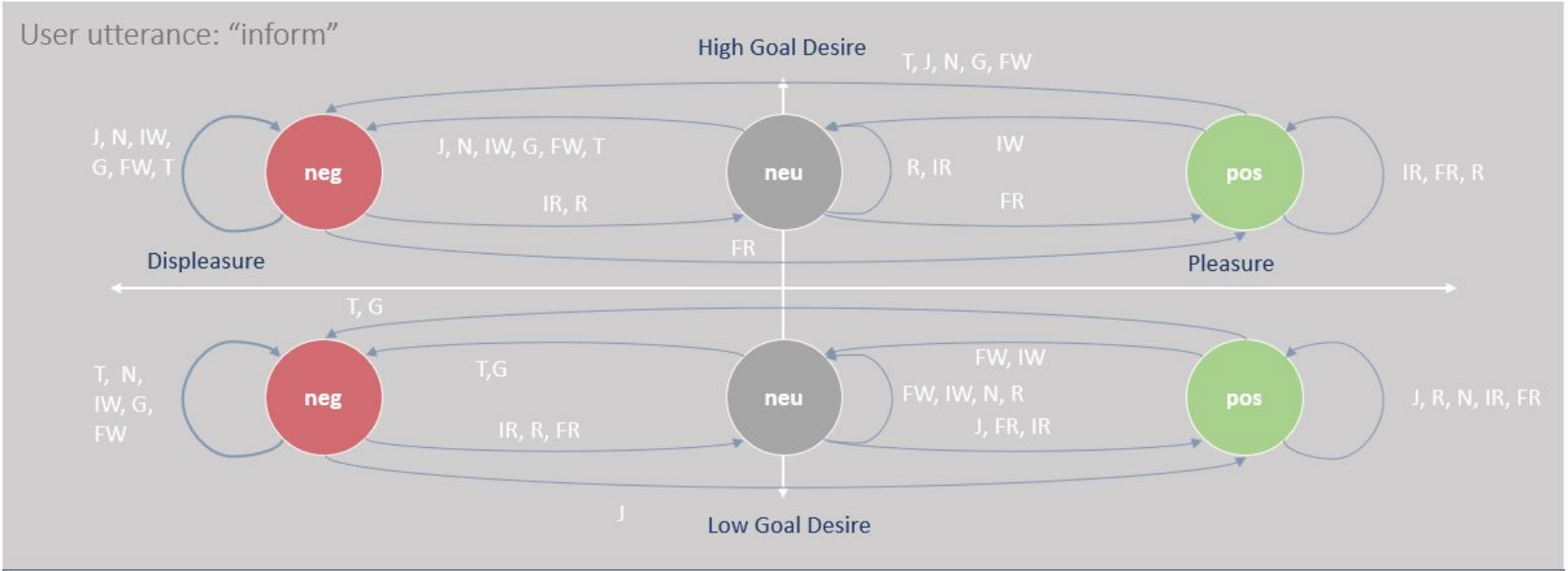
Concept Intrinsic Motivation

User Mood Transition Model



Concept Mood Based IM

User Mood Transition Model



R: utter_request, **J**: joke, **N**: utter_nothing, **IR**: inform_right, **IW**: inform_wrong, **IR**: inform_right, **G**: utter_goodbye, **T**: trigger_user, **FR**: find_drink_right, **FW**: find_drink_wrong

Concept Mood Based IM

Intrinsic Reward - Reward Formulation

Transition Reward:

$$r_{transition}(em_{t+1} \mid em_t) = \begin{Bmatrix} r_{neg-neg} & r_{neg-neu} & r_{neg-pos} \\ r_{neu-neg} & r_{neu-neu} & r_{neu-pos} \\ r_{pos-neg} & r_{pos-neu} & r_{pos-pos} \end{Bmatrix}$$

Mood Reward:

$$r_{mood} = r_{transition} + \begin{cases} 0 & if\ goaldesire = high \\ r_{goaldesire} & else \end{cases} = r_{inrinsic}$$

Extrinsic Reward:

$$r_{ext} = r_{step} + \begin{cases} -w_{fail} \cdot \max steps & if\ goal\ was\ not\ found \\ w_{success} \cdot \max steps & else \end{cases}$$

Mood Reward:

$$r_{overall} = \beta_{reward} \cdot r_{intrinsic} + (1 - \beta_{reward}) \cdot r_{ext}$$

Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

Concept Curiosity Driven IM

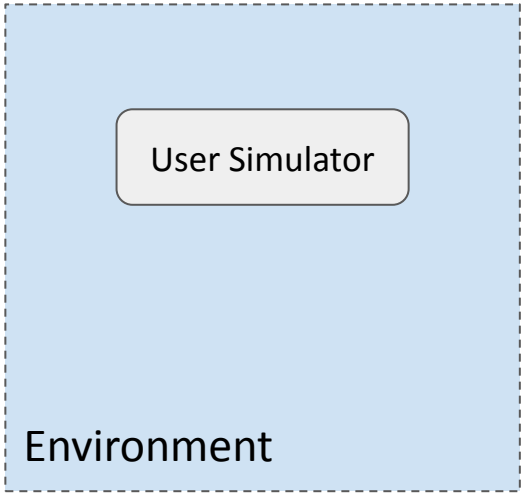
Motivation

1. **Humans don’t learn skills randomly but by curiosity**
2. **Temporal and complexity ordered learning**
3. **Spent more time on complexer tasks**
4. **Our agent should incorporate curiosity instead of e-greedy exploration**

Concept Curiosity Driven IM

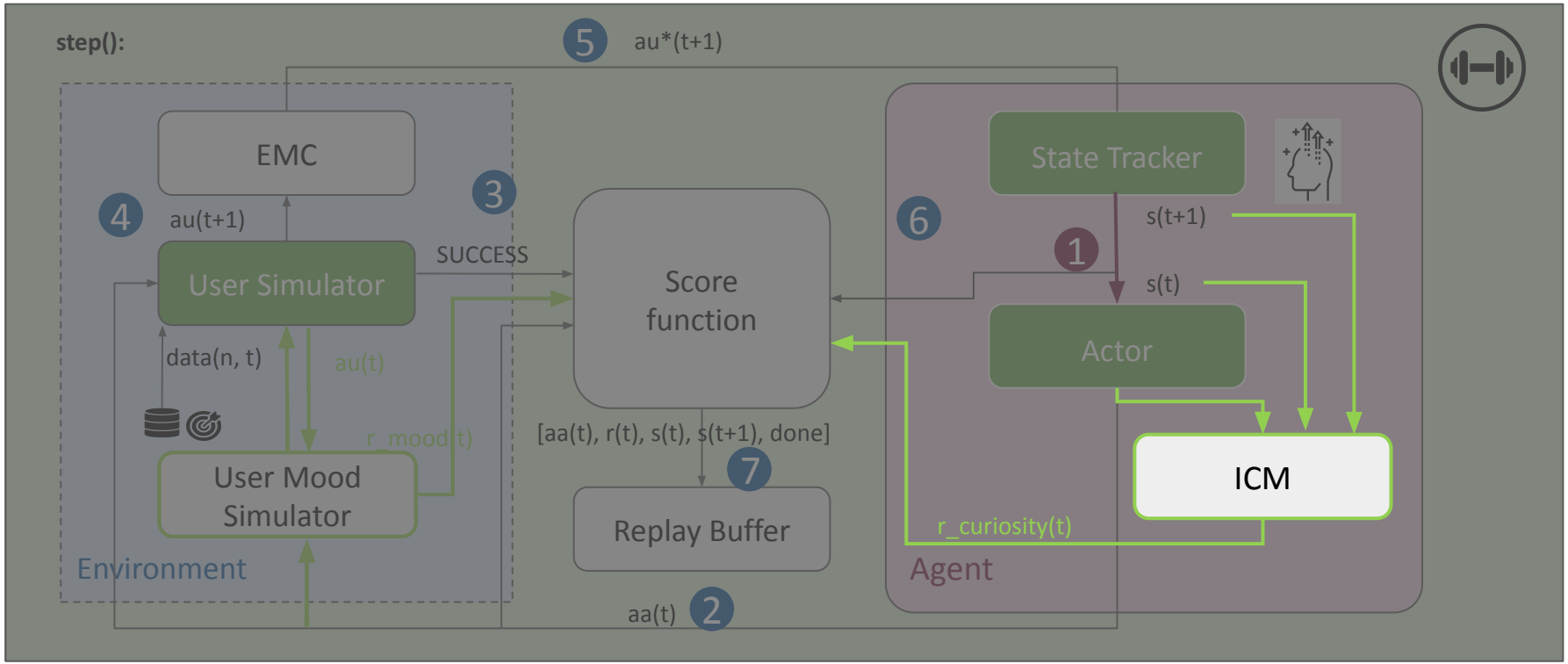
Adaption of the Agent

Goal



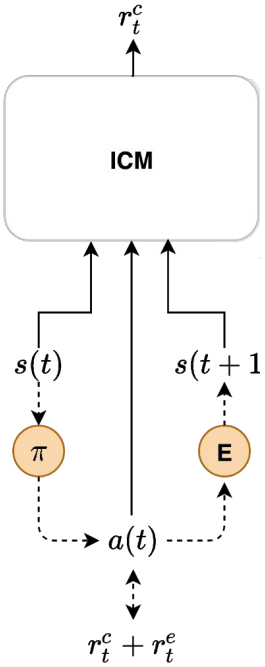
Concept Curiosity Driven IM

Adaption of the Agent



Concept Curiosity Driven IM

Implementation

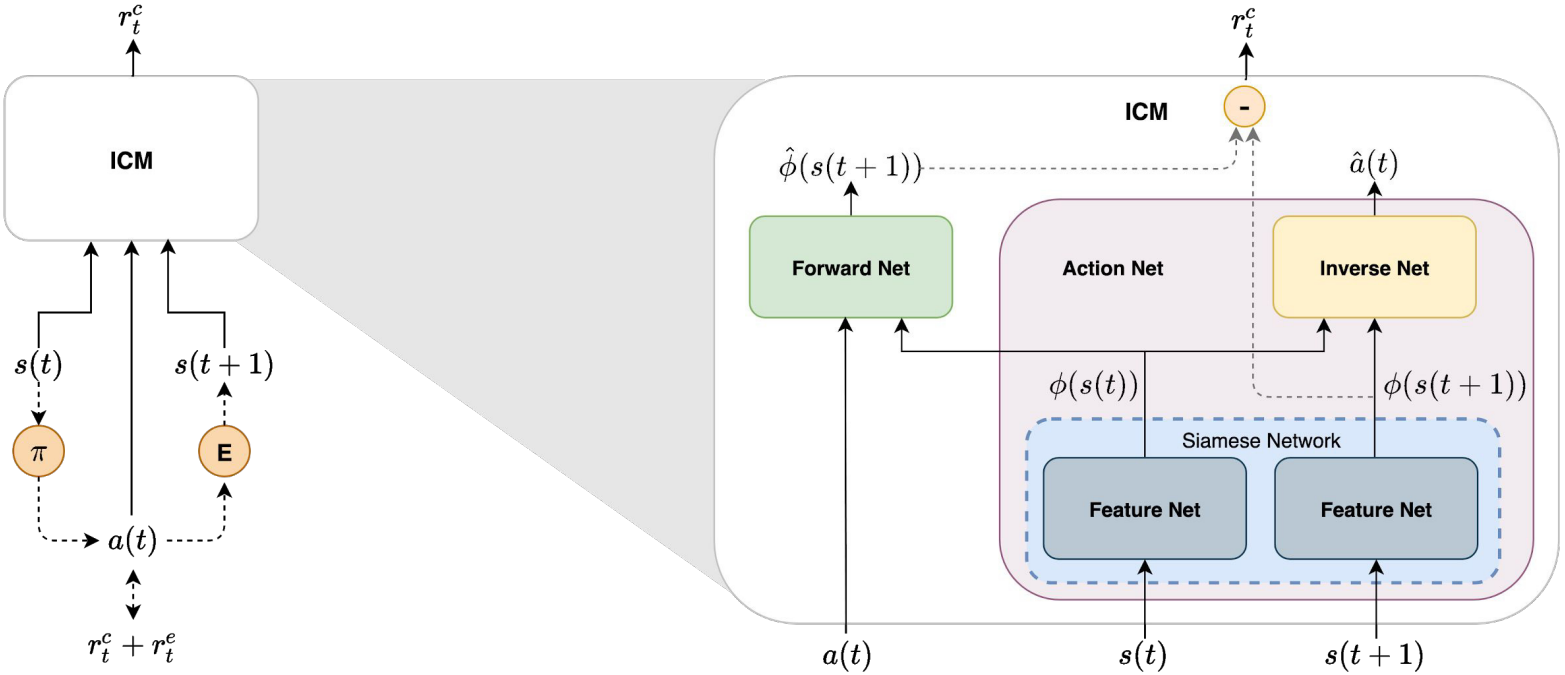


1. Get sample s(t), a(t) and s(t+1)

2. Get reward that encodes informativity of a(t) to get from s(t) to s(t+1)

Concept Curiosity Driven IM

Implementation



Concept Curiosity Driven IM

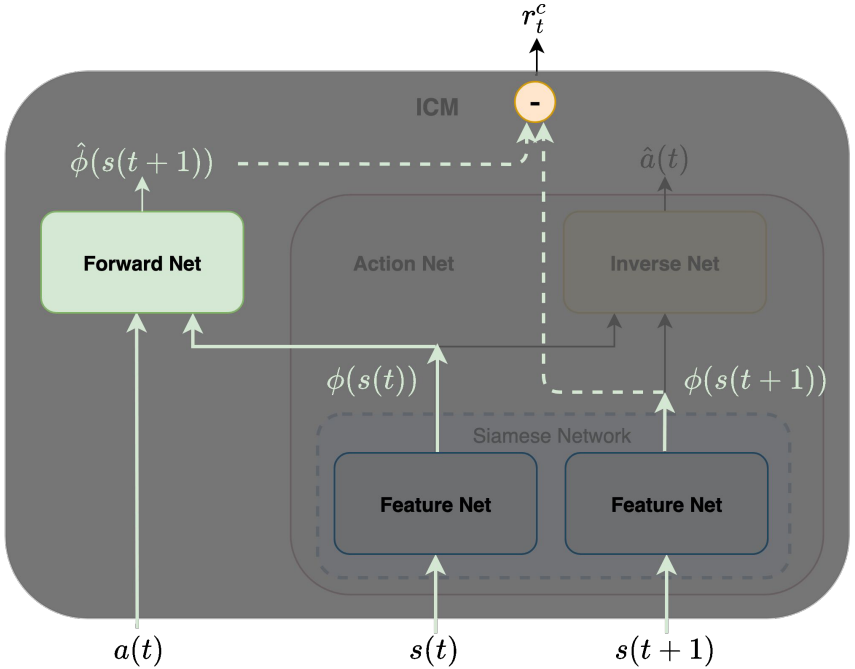
Implementation

Curiosity Reward:

$$r_c^t = ||\hat{\phi}(s(t+1)) - \phi(s(t+1))||^2$$

Loss Formulation Forward

$$L_{forward} = ||\hat{\phi} - \phi||^2$$

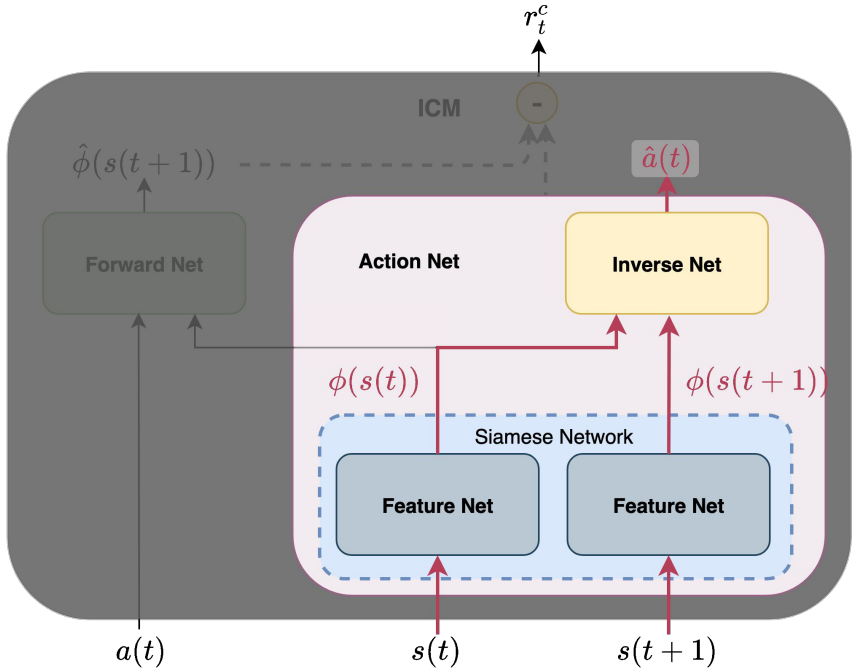


Concept Curiosity Driven IM

Implementation

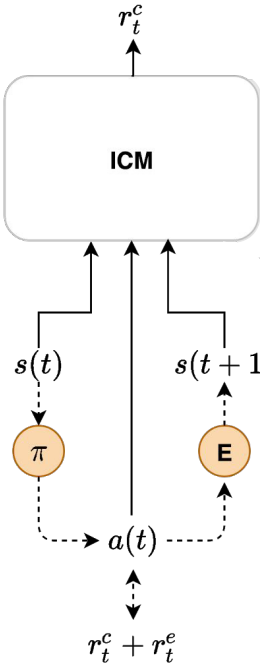
Loss Formulation Action

$$L_{action} = CCE(\hat{a}(t), a(t))$$



Concept Curiosity Driven IM

Overall Reward Formulation



Intrinisic Reward:

$$r_{intrinsic} = \alpha_{reward} \cdot r_{mood} + (1 - \alpha_{reward}) \cdot r_c$$

Overall Reward:

$$r_{overall} = \beta_{reward} \cdot r_{intrinsic} + (1 - \beta_{reward}) \cdot r_{ext}$$

Overall Optimization

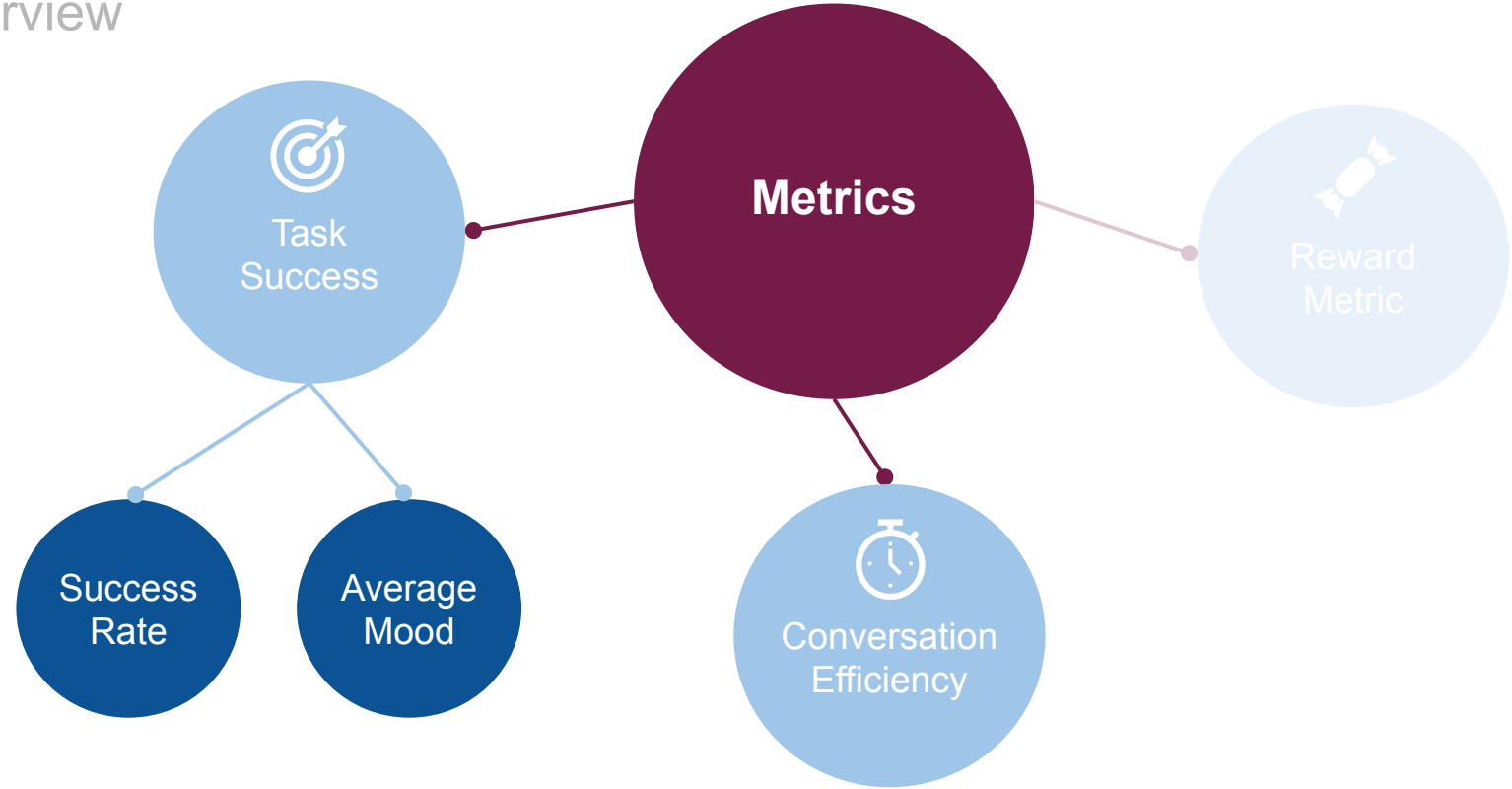
$$\min \; - L_{agent}(r_{overall}) + \beta_{cur} \cdot L_{forward} + (1 - \beta_{cur}) \cdot L_{action}$$

Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

Metrics

Overview



Metrics

Task Success

Success Rate:

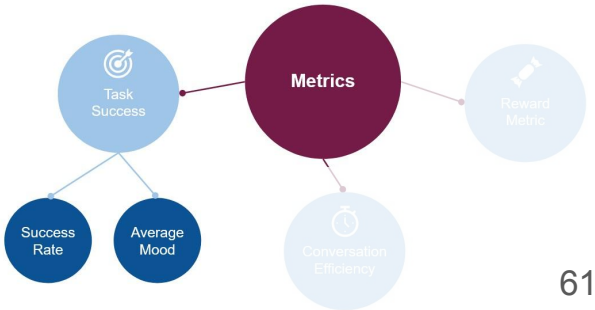
$$M_{success} = \frac{|\ conversations\ in\ which\ the\ user's\ goal\ is\ met\ |}{|\ all\ conversations\ |}$$

Average Mood:

$$M_{mood} = \frac{\Sigma_{i=1}^{max\_round} mood_{user}(i)}{max\_round}$$

where

$$mood_{user} = \begin{cases} 0 & \text{if mood is negative} \\ 0.5 & \text{if mood is neutral} \\ 1 & \text{if mood is positive} \end{cases}$$



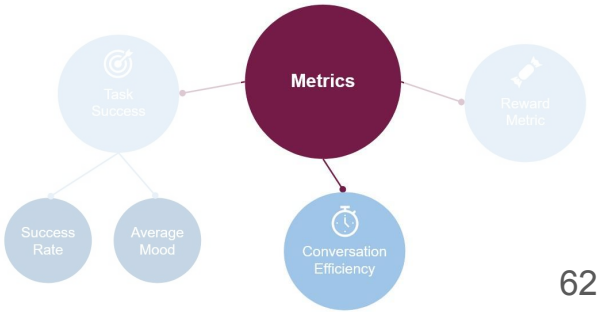
Metrics

Conversation Efficiency

$$M_{eff} = \begin{cases} \frac{|unique\ agent\ actions|}{|agent\ actions|}, & \text{if goal desire = high} \\ \frac{|unique\ agent\ actions|}{|agent\ actions^*|}, & \text{if goal desire = low} \end{cases}$$

where $|agent\ actions|$ is the number of actions taken by the agent

and $|agent\ actions^*|$ is the number of actions taken by the agent counting the actions *joke* and *utter_nothing* only once.

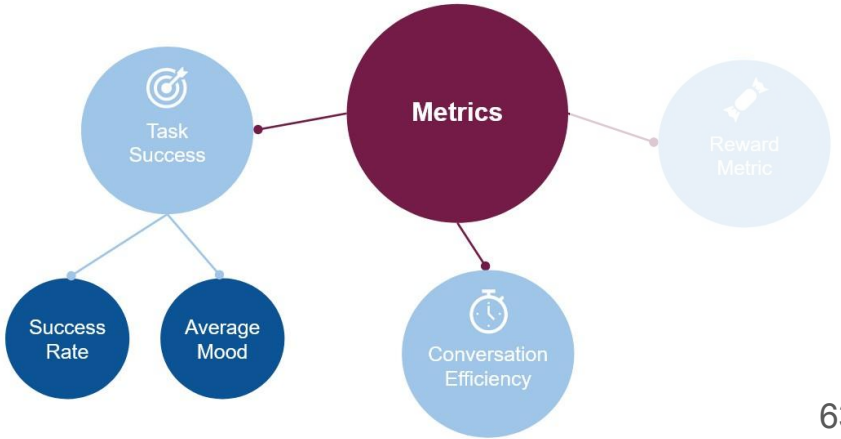


Metrics

Quality Metric

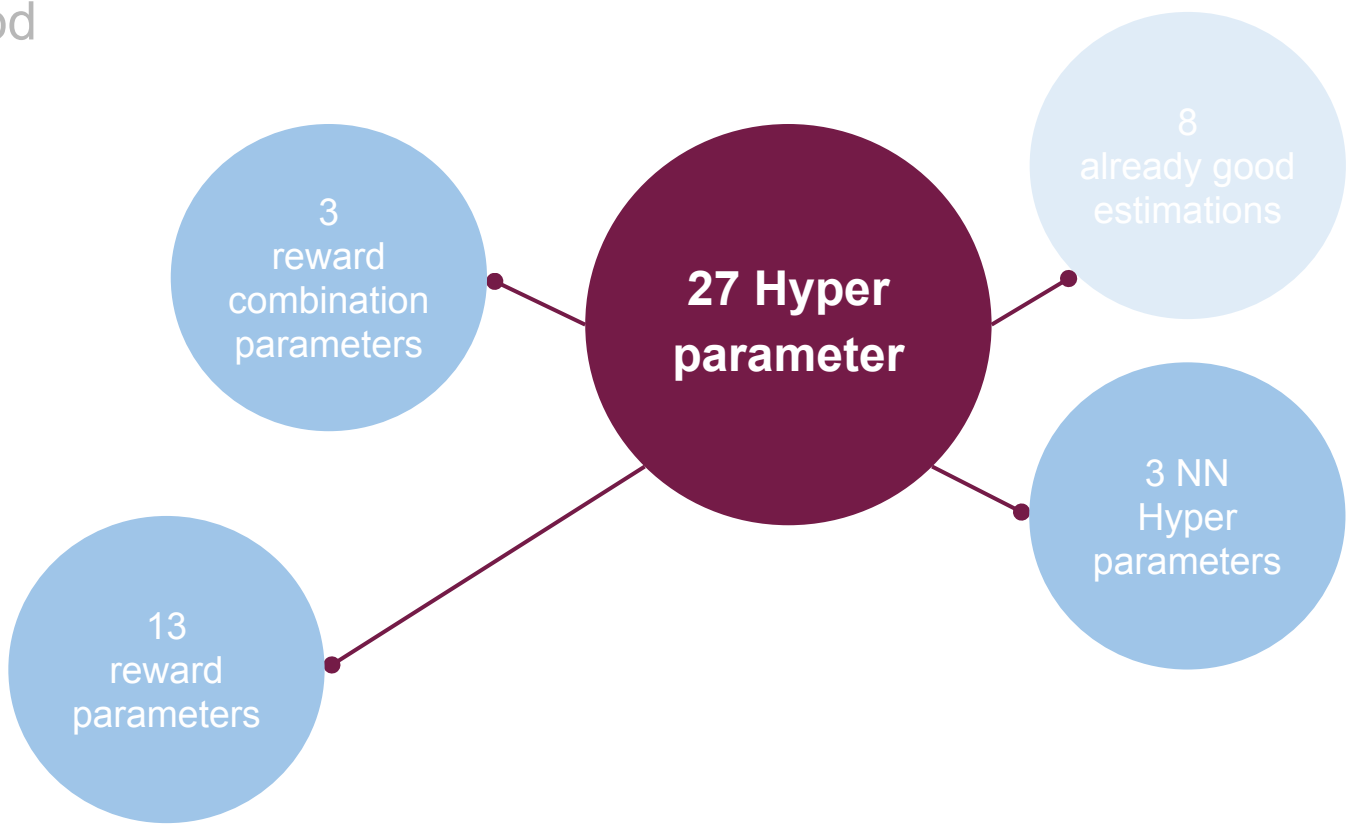
$$M_{quality} = \alpha_{met} M_{eff} + \gamma_{met} M_{mood} + \delta_{met} M_{success}$$

weighted sum of all presented metrics



Hyperparameter Tuning

Method

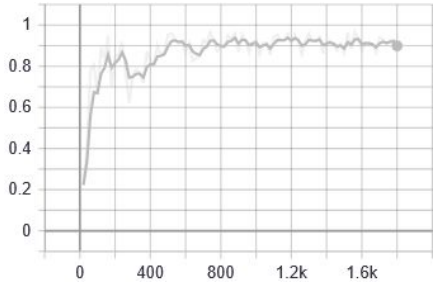


Validation Motivation Concepts

Results Metrics

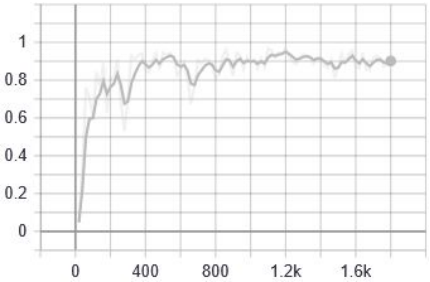
All Rewards

Quality metric



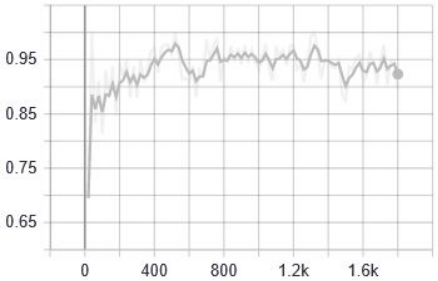
=

Average mood



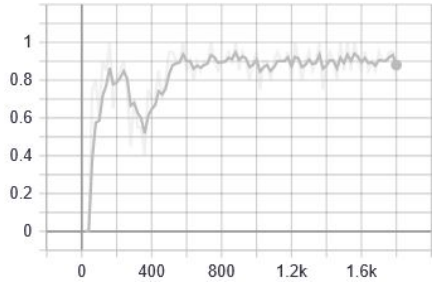
+

Average efficiency



+

Success rate



Hyperparameter Tuning

Example High Goal Desire

Inform: cold **Request: SIZE, DRINK**

Can I have a cold drink Which SIZE do you have?

We have medium

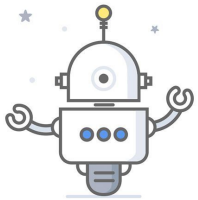
Which DRINK do you have?

We have the following drink: coffee medium cold 15

Thanks. I will take this :)

bye bye

goodbye



Hyperparameter Tuning

Example Low Goal Desire

Inform: cold **Request: SIZE, DRINK**

...

Hey, sorry to disturb you. What DRINK can I offer to you?

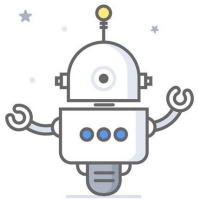
Which DRINK do you have?

We have medium

Which DRINK do you have?

Helvetica and Times New Roman walk into a bar. Get out of here! shouts the bartender. We don't serve your type.

...



Hyperparameter Tuning

Example Low Goal Desire

Inform: cold **Request: SIZE, DRINK**

We have the following drink: green tea medium hot 32

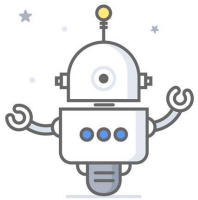
No thats not what I wanted 😡

We have red vine

I would like this to be cold 🤔

...

... 🤔



Hyperparameter Tuning

Example Low Goal Desire

Inform: cold **Request: SIZE, DRINK**

...

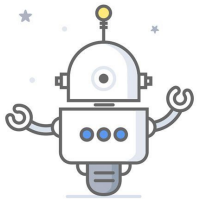
... 🤔

We have the following drink: red vine Jumbo cold 35

Thanks. I will take this :) 😄

bye bye

goodbye 😄



Validation Motivation Concepts

Experiment

Ablation study to test, the contribution of the Intrinsic Motivation (IM) concepts

1 Only Extrinsic Motivation

$$r_{overall} = r_{ext}$$

Pure extrinsic motivation reward, similar to the baseline agent

2 Only Extrinsic & Mood Based IM

$$r_{overall} = r_{ext} + r_{mood}$$

No curiosity Driven IM reward, to test the contribution of it

3 Only Extrinsic & Curiosity Driven IM

$$r_{overall} = r_{ext} + r_{curiosity}$$

No Mood Based IM reward, to test the contribution of it

4 Only Intrinsic Motivation

$$r_{overall} = r_{mood} + r_{curiosity}$$

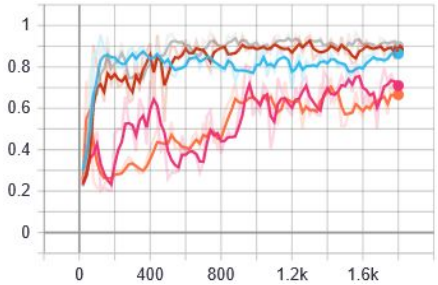
Only mood and curiosity rewards, to test how important the extrinsic reward is

Validation Motivation Concepts

Results Metrics

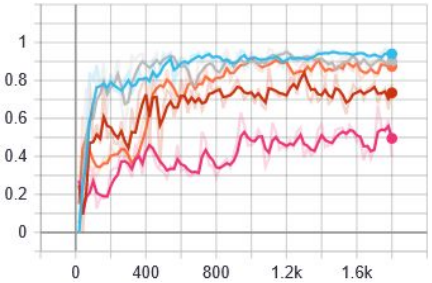
All Rewards
Only Extrinsic
Extrinsic + Mood Reward
Extrinsic + Curiosity Reward
Curiosity + Mood Reward

Quality metric



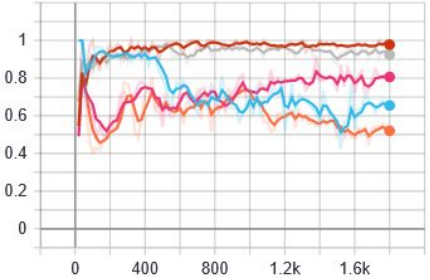
=

Average mood



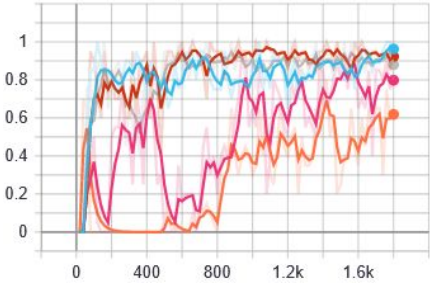
+

Average efficiency



+

Success rate



Validation Motivation Concepts

Experiment

1 Only Extrinsic Motivation

High Goal Desire
short & efficient conversation

Low Goal Desire
no jokes or long conversations anymore

› no goal desire distinguishing

2 Only Extrinsic & Mood Based IM

High Goal Desire
short & efficient conversation

Low Goal Desire
Many repetitions leading to a bad conversation flow

› goal desire distinguishing

3 Only Extrinsic & Curiosity Driven IM

High Goal Desire
no efficient but creative answer

Low Goal Desire
not finishing the conversation

› goal desire distinguishing

4 Only Intrinsic Motivation

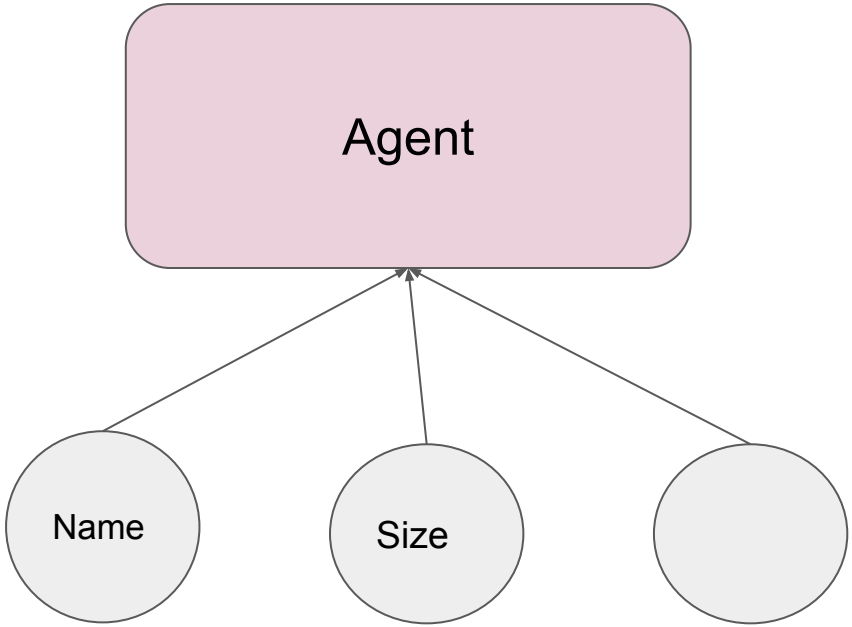
High Goal Desire
no efficiency, not finishing

Low Goal Desire
No goal finding at all
Long conversations with only “jokes” and “nothing”

› no goal desire distinguishing

Continuous Skill Expansion

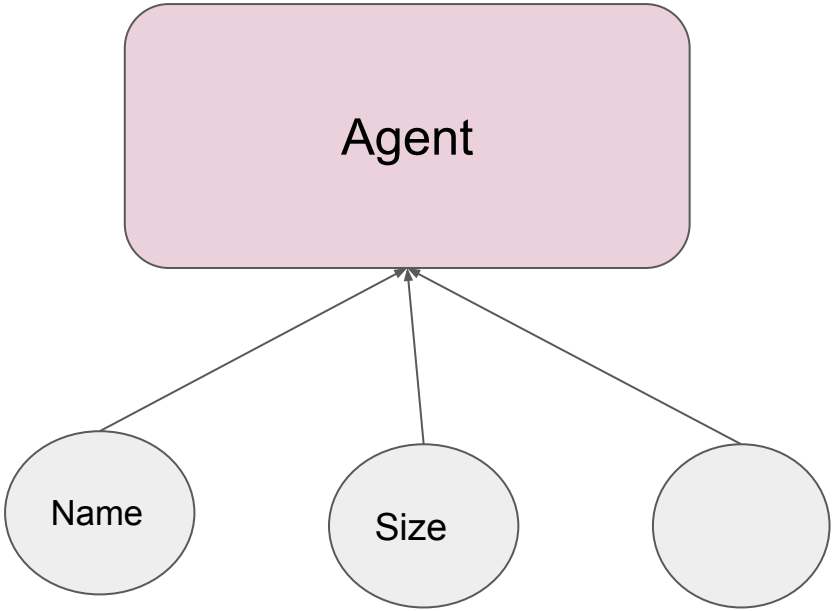
Experiment



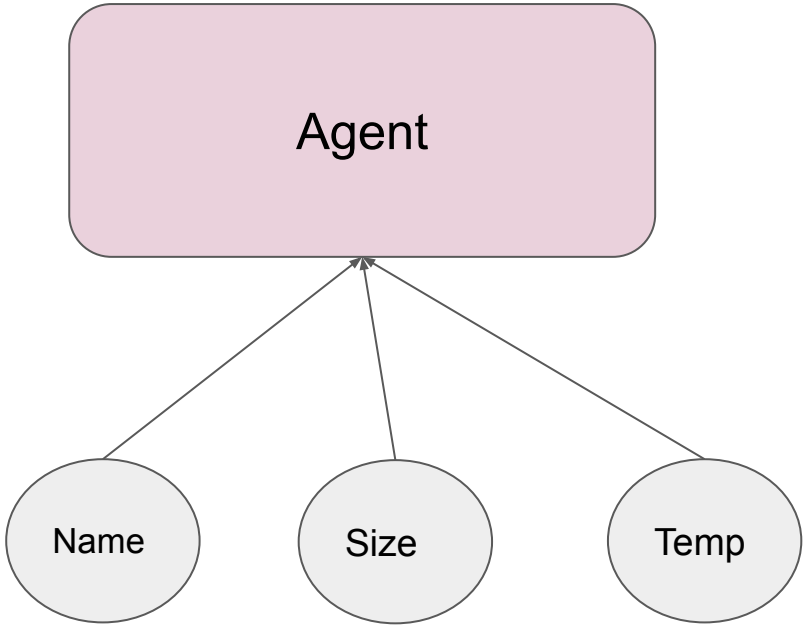
User stories at training time

Continuous Skill Expansion

Experiment



User stories at training time



User stories at test time

Continuous Skill Expansion

Results

```
Agent Action: Which DRINK do you like?
***User Response: I would like this to be white vine
 ---User mood:neutral
Agent Action: Which SIZE do you like?
***User Response: Which SIZE do you have?
 ---User mood:negativ
Agent Action: Which TEMP do you like?
***User Response: I would like this to be cold
 ---User mood:negativ
Agent Action: We have the following drink: white vine small cold 39
***User Response: Thanks. I will take this :)
 ---User mood:positiv
Agent Action: bye bye
***User Response: goodbye
 ---User mood:positiv
```

Correct behaviour

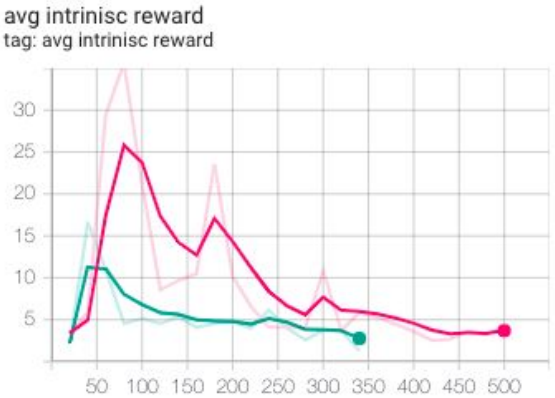
```
Initial User Utterance: Can I have a medium drink
Agent Action: Which DRINK do you like?
***User Response: I would like this to be barracuda
 ---User mood:neutral
Agent Action: We have the following drink: barracuda medium cold 28
***User Response: Thanks. I will take this :)
 ---User mood:positiv
Agent Action: bye bye
***User Response: goodbye
 ---User mood:positiv
```

Actual behaviour

Continuous Skill Expansion

Results

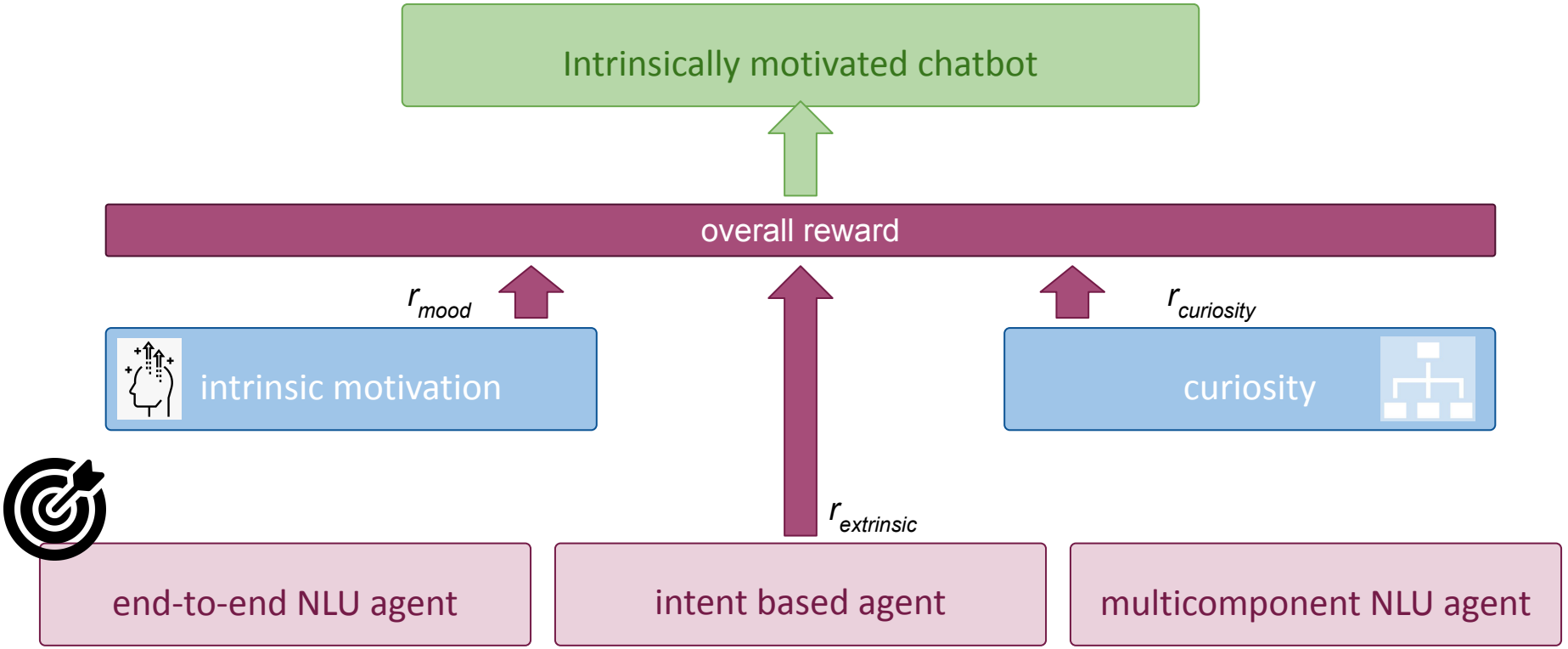
	Number of episodes
Agent (Partially filled slots)	260
Agent (Scratch)	500



Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

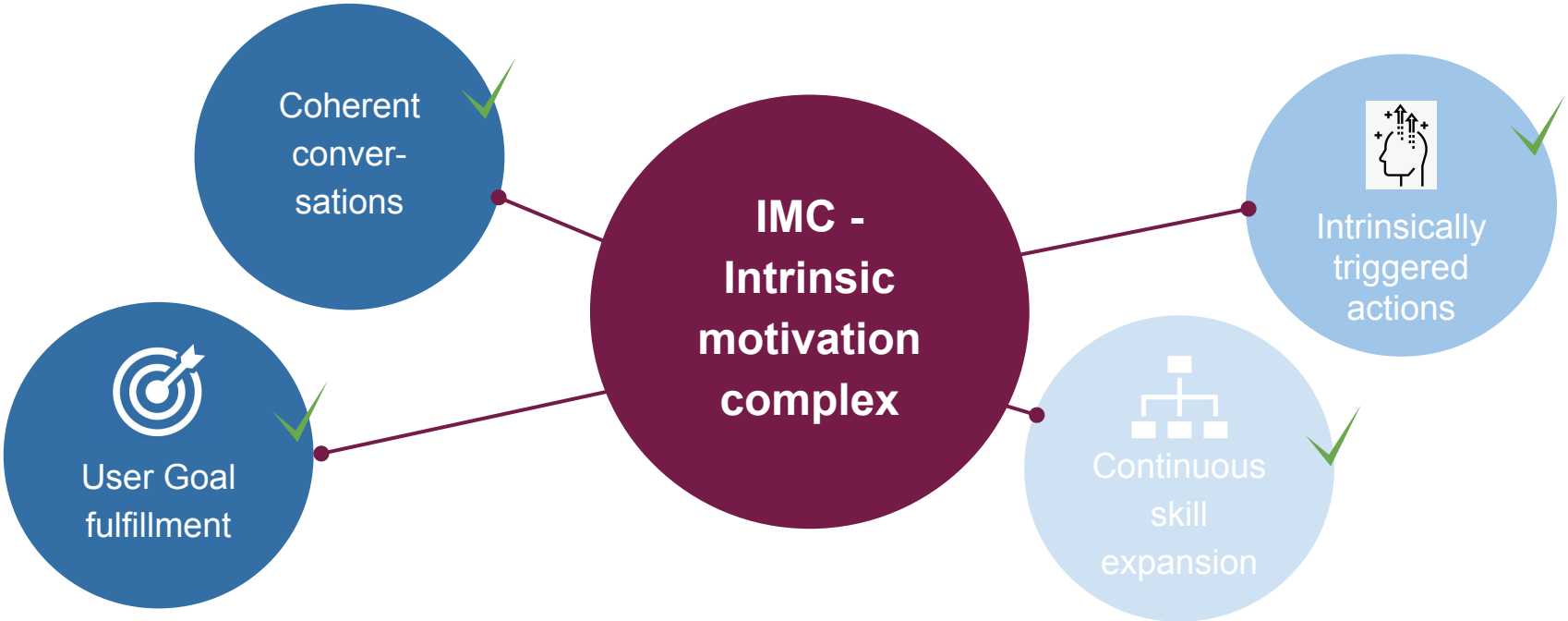
Conclusion


Intrinsically motivated chatbot
overall reward
r_{mood}
$r_{curiosity}$
intrinsic motivation
curiosity
$r_{extrinsic}$
end-to-end NLU agent
intent based agent
multicomponent NLU agent

Project Overview

Overall Objectives

Intrinsic motivation complex for an artificial conversational assistant



Agenda

1. Project Motivation and Overview
2. Scientific Concepts
3. Baseline Agents
4. Intrinsic Motivation
5. Experiments
6. Conclusion
7. Demonstration

Sources and Literature

Sutton, Richard S.; Barto, Andrew G. (2018). Reinforcement Learning: An Introduction (2 ed.). MIT Press. ISBN 978-0-262-03924-6.

https://dictionary.apa.org/intrinsic-motivation

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, Trevor Darrell (2017). Curiosity-driven Exploration by Self-supervised Prediction. https://arxiv.org/abs/1705.05363.

Karol Gregor, Danilo Jimenez Rezende, Daan Wierstra (2016). Variational Intrinsic Control. https://arxiv.org/abs/1611.07507.

M. Brenner (2018). Training a goal oriented chatbot. https://towardsdatascience.com/training-a-goal-oriented-chatbot-with-deep-reinforcement-learning-part-i-introduction-and-dce3af21d383.

Andrew Y. Ng., Daishi Harada, Stuart Russel (1999). Policy invariance under reward transformations: Theory and application to reward shaping. https://arxiv.org/pdf/1908.06976.pdf

A. Aubret, L. Matignon, S. Hassas (2019). A survey on intrinsic motivation in reinforcement learning. http://www.robotics.stanford.edu/~ang/papers/shaping-icml99.pdf