Compositional Instruction Following with Language Models and Reinforcement Learning

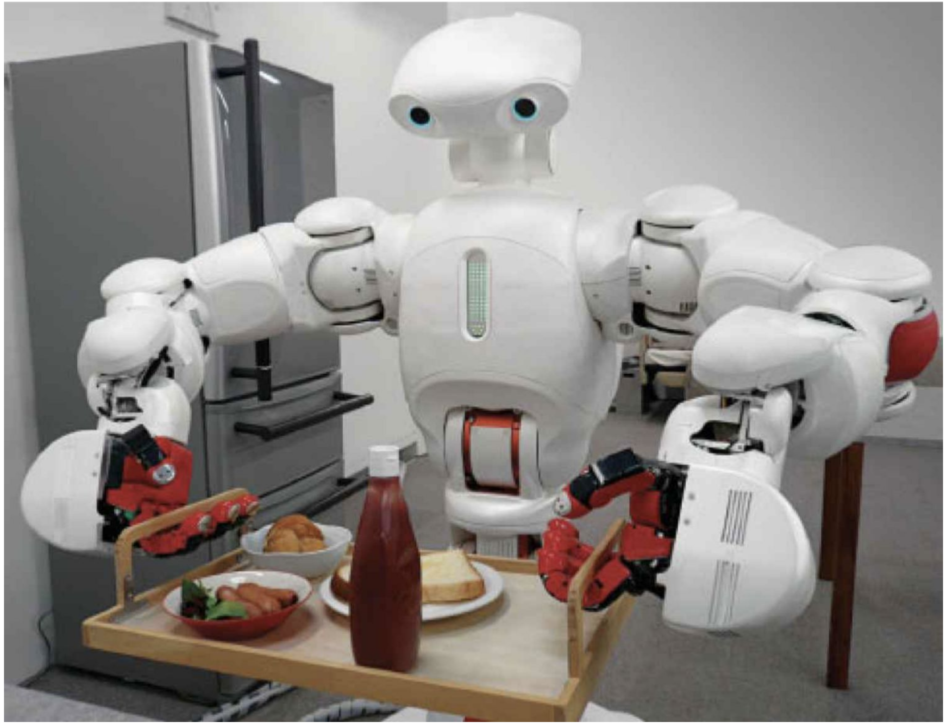
Vanya Cohen*, **Geraud Nangue Tasse***,
Nakul Gopalan, Steven James, Matthew Gombolay, Ray Mooney, Benjamin Rosman

RLC 2025

TEXAS The University of Texas at Austin
WITS UNIVERSITY
ASU
Georgia Tech
TRANSACTIONS tmlr on ML RESEARCH

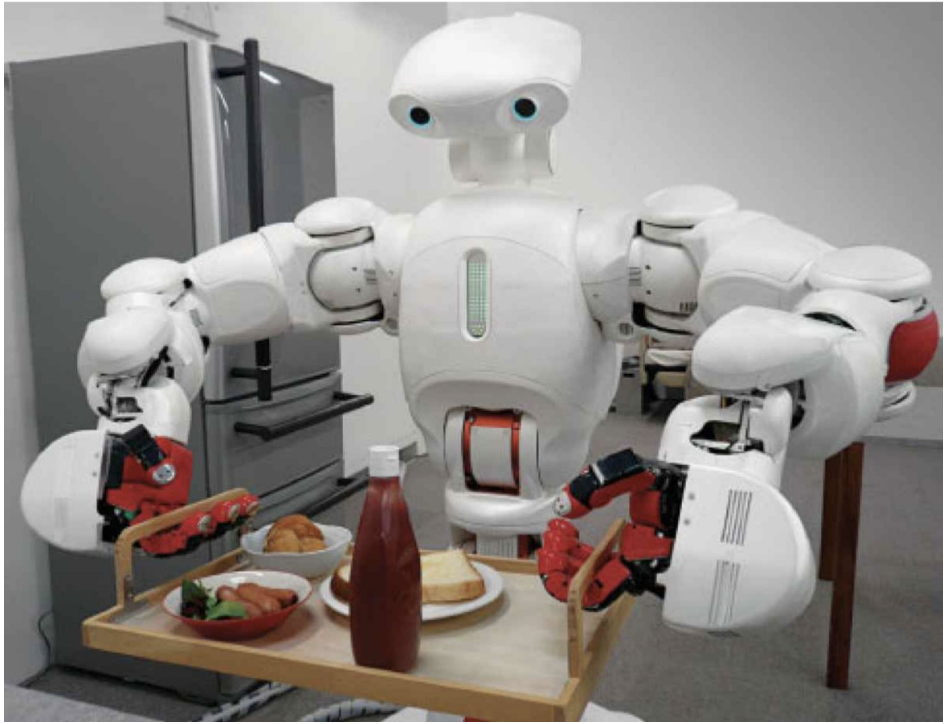
Language and RL Tasks Share Compositional Structure

- “**Serve breakfast** with **plain toast** *and* **ketchup**…”



Language and RL Tasks Share Compositional Structure

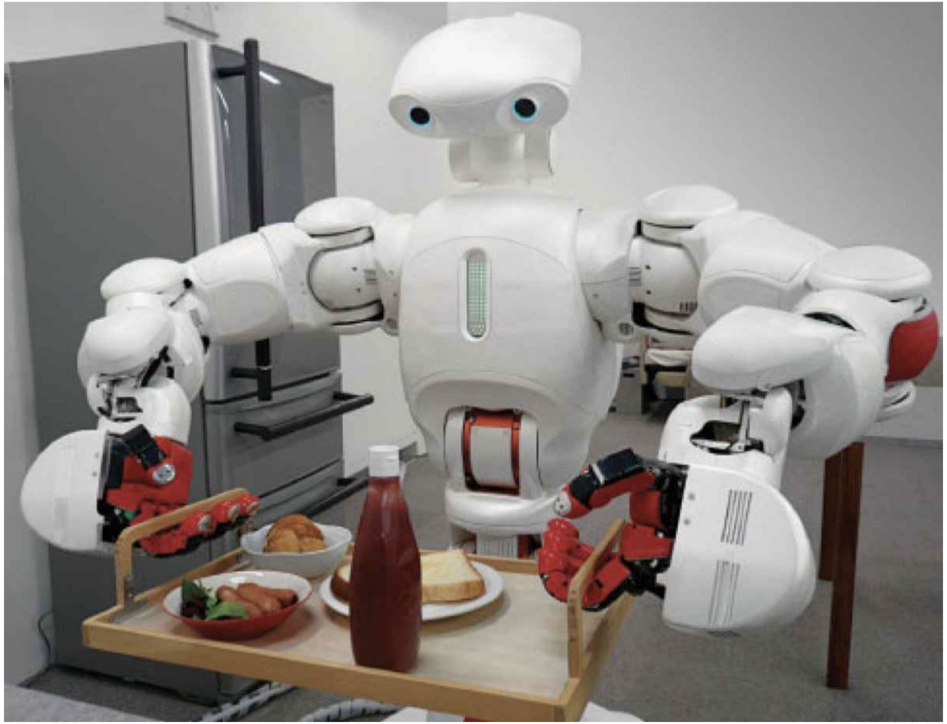
- "**Serve breakfast** with **plain toast** *and* **ketchup**…"
- Neural networks **struggle to generalize compositionally**[1].



1. Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. Proceedings of the 35th International Conference on Machine Learning.

Language and RL Tasks Share Compositional Structure

- "**Serve breakfast** with **plain toast** *and* **ketchup**…"
- Neural networks **struggle to generalize compositionally**[1].
- Compose existing policies to perform tasks with minimal training.



1. Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. Proceedings of the 35th International Conference on Machine Learning.

World Value Functions (Tasse et al. 2020, 2022)

$$Q_{\pi}(s, a) = \mathbb{E}_{s}^{\pi}\left[\sum_{t=0}^{\infty} \bar{r}(s_t, a_t)\right]$$

1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

World Value Functions (Tasse et al. 2020, 2022)

- Add a goal *g* to the Q function.
- WVF represents how to achieve all goals **and** their value
- Learn one WVF for each task in the environment we wish to compose.

$$Q_\pi(s, g, a) = \mathbb{E}_s^\pi\left[\sum_{t=0}^{\infty} \bar{r}(s_t, g, a_t)\right]$$

1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

World Value Functions (Tasse et al. 2020, 2022)

- Add a goal *g* to the Q function.
- WVF represents how to achieve all goals **and** their value
- Learn one WVF for each task in the environment we wish to compose.
- <u>Train by **penalizing** the agent for entering a terminal state for another goal.</u>

$$Q_\pi(s, g, a) = \mathbb{E}_s^\pi\left[\sum_{t=0}^{\infty} \bar{r}(s_t, g, a_t)\right]$$

$$\bar{r}(s, g, a) = \begin{cases} \bar{r}_{MIN} & if\ g \neq s \in G \\ r(s, a) & otherwise \end{cases}$$

1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

World Value Functions (WVF) (Tasse et al. 2020, 2022)

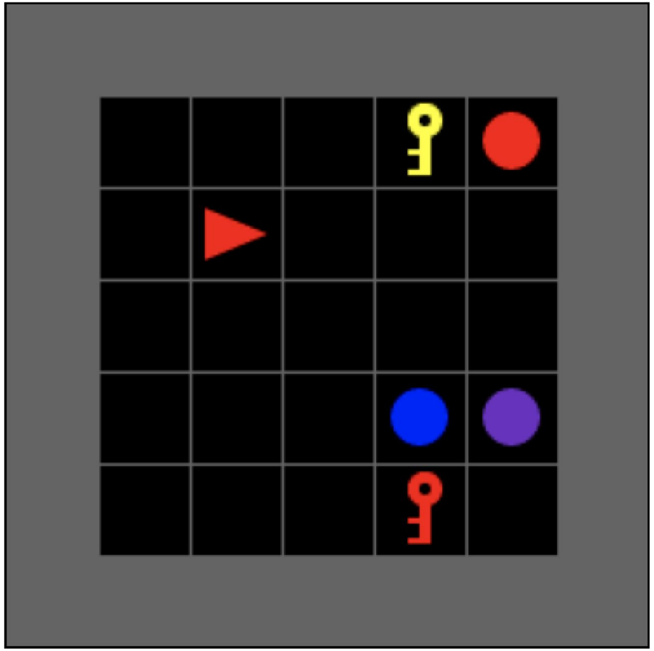
- Compose these WVFs.
    - Arbitrary expressions of **AND**, **OR**, and **NOT**.
    - Can now solve a **combinatorial number** of goal reaching tasks

For AND (conjunction) the composed WVF is given by:

$$\begin{aligned} \bar{Q}_1^* \wedge \bar{Q}_2^* :\ & \mathcal{S} \times \mathcal{G} \times \mathcal{A} \to \mathbb{R} \\ & (s, g, a) \mapsto \min\{\bar{Q}_1^*(s, g, a), \bar{Q}_2^*(s, g, a)\} \end{aligned}$$

BabyAI (Chevalier-Boisvert et al. 2019)

- Gridworld domain consisting of navigation tasks.

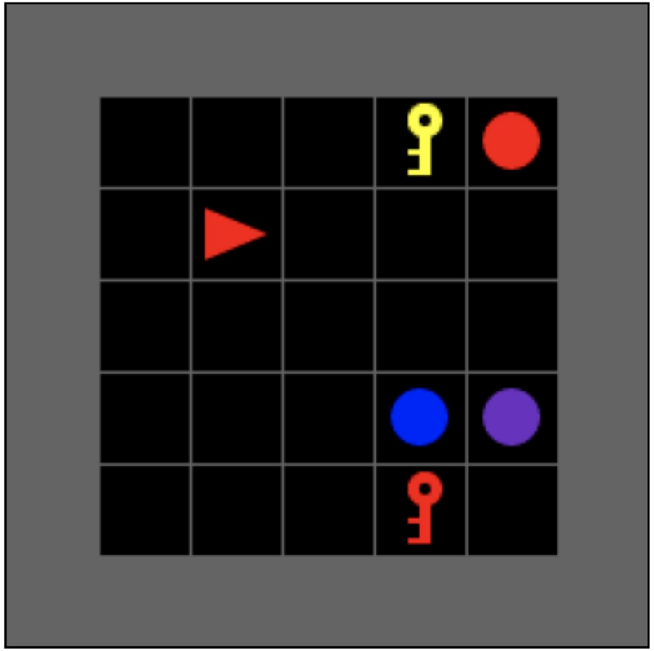


BabyAI (Chevalier-Boisvert et al. 2019)

- Gridworld domain consisting of navigation tasks.

“Pick up a red object” $\wedge$ “Pick up a key”

$\neg$“Pick up a blue object” $\vee$ “Pick up the ball”



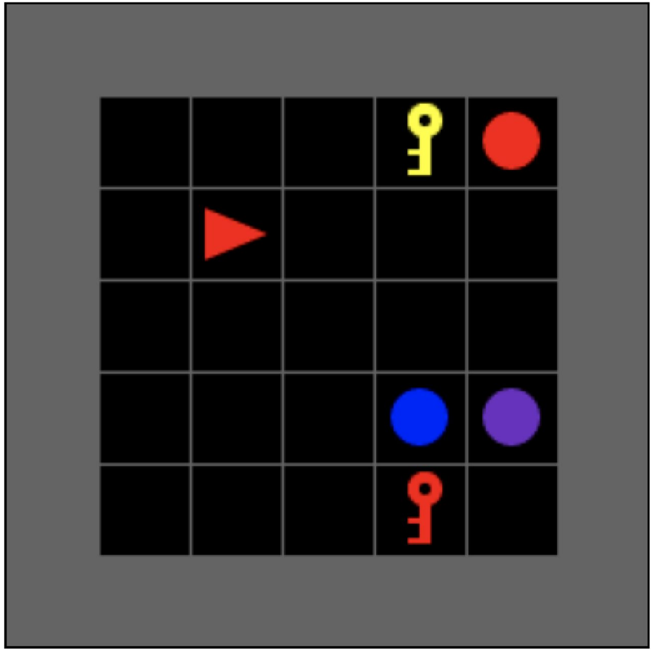
BabyAI (Chevalier-Boisvert et al. 2019)

- Gridworld domain consisting of navigation tasks.

“Pick up a red object” $\wedge$ “Pick up a key”

$\neg$“Pick up a blue object” $\vee$ “Pick up the ball”

- Modified task set to include 162 goal reaching tasks that can be solved through **AND**, **OR**, and **NOT** expressions over object attributes.



Compositionally-Enabled RL and Language Agent (CERLLA)

- How can we use compositionality of language + value functions to generalize better?

Compositionally-Enabled RL and Language Agent (CERLLA)

- How can we use compositionality of language + value functions to generalise better?
- Must learn mapping from natural language to WVF composition

Compositionally-Enabled RL and Language Agent (CERLLA)

- How can we use compositionality of language + value functions to generalise better?
- Must learn mapping from natural language to WVF composition.
- Idea: Use language models to translate instruction into formal language / boolean symbols (e.g. semantic parsing).

Compositionally-Enabled RL and Language Agent (CERLLA)

- How can we use compositionality of language + value functions to generalise better?
- Must learn mapping from natural language to WVF composition.
- Idea: Use language models to translate instruction into formal language / boolean symbols (e.g. semantic parsing).
- But these symbols are arbitrary (just an index over WVFs) - how do we know translation is correct?

Compositionally-Enabled RL and Language Agent (CERLLA)

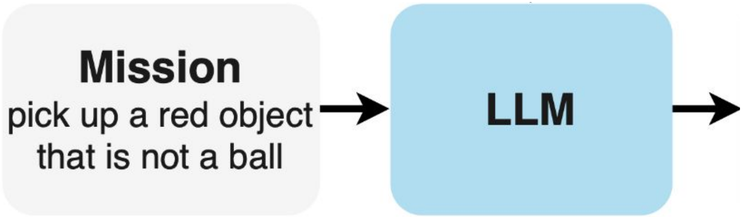
- How can we use compositionality of language + value functions to generalise better?
- Must learn mapping from natural language to WVF composition
- Idea: Use language models to translate instruction into formal language / boolean symbols (e.g. semantic parsing).
- But these symbols are arbitrary (just an index over WVFs) - how do we know translation is correct?
- Idea: Use environment feedback to learn the translation!

Compositionally-Enabled RL and Language Agent (CERLLA)

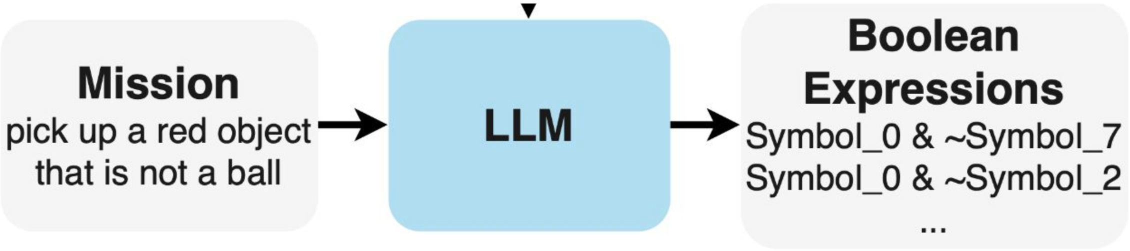
- How can we use compositionality of language + value functions to generalise better?
- Must learn mapping from natural language to WVF composition
- Idea: Use language models to translate instruction into formal language / boolean symbols (e.g. semantic parsing).
- But these symbols are arbitrary (just an index over WVFs) - how do we know translation is correct?
- Idea: Use environment feedback to learn the translation!

Core challenge: CERLLA learns to parse input commands to **arbitrary symbols** representing WVFs with **unknown semantics**, using **environment rollouts**, a much noisier form of supervision than is typical for weakly supervised parsing methods.

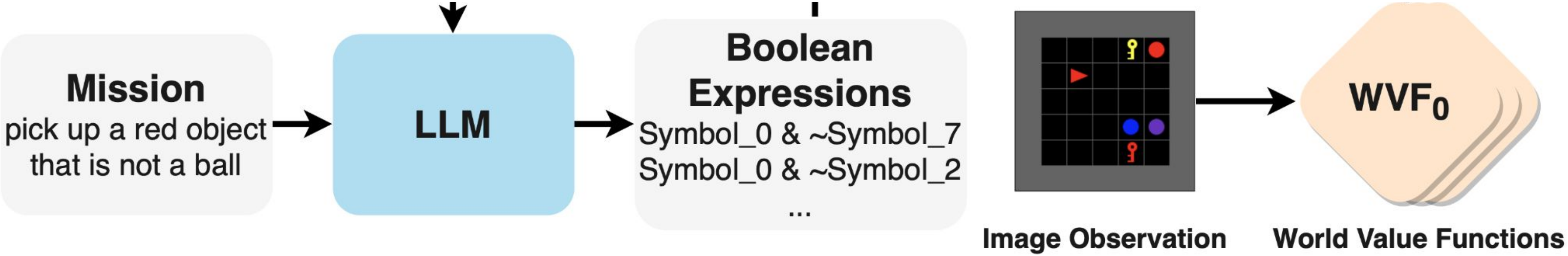
Compositionally-Enabled RL and Language Agent (CERLLA)



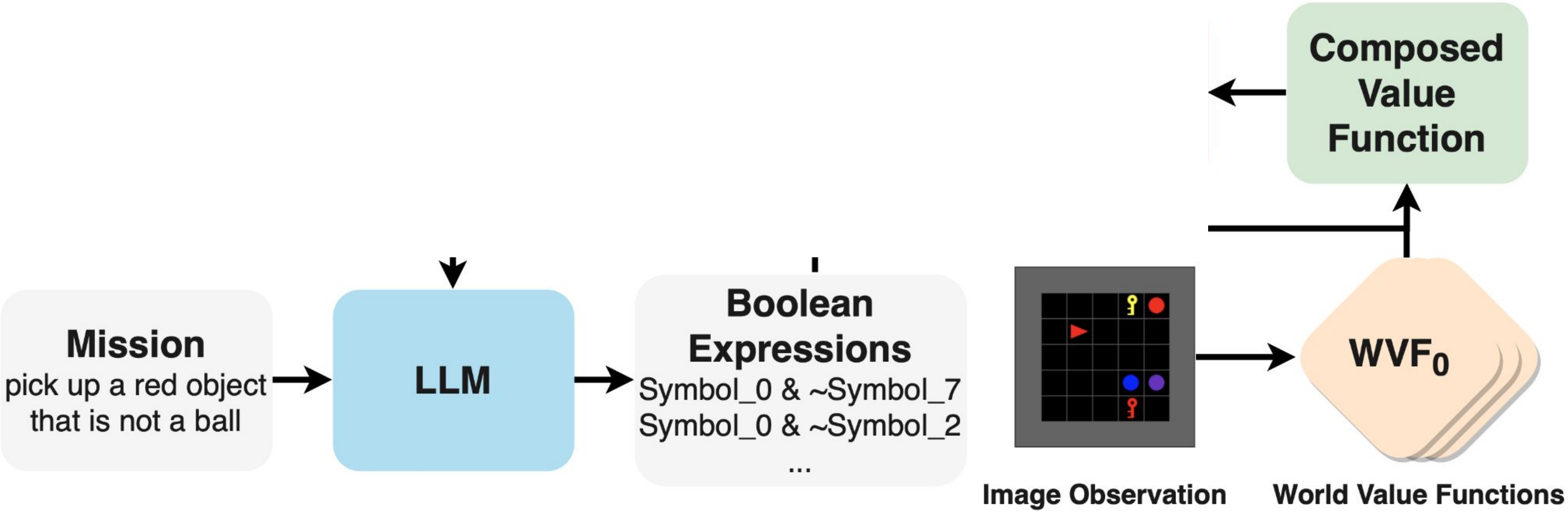
Compositionally-Enabled RL and Language Agent (CERLLA)



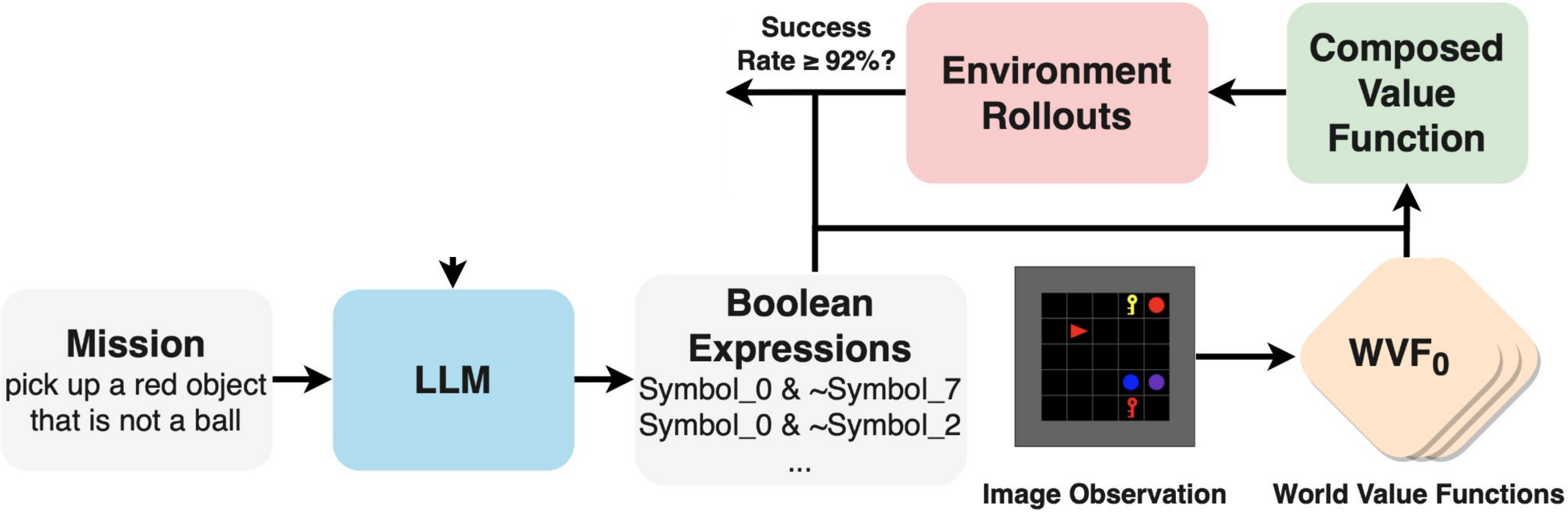
Compositionally-Enabled RL and Language Agent (CERLLA)



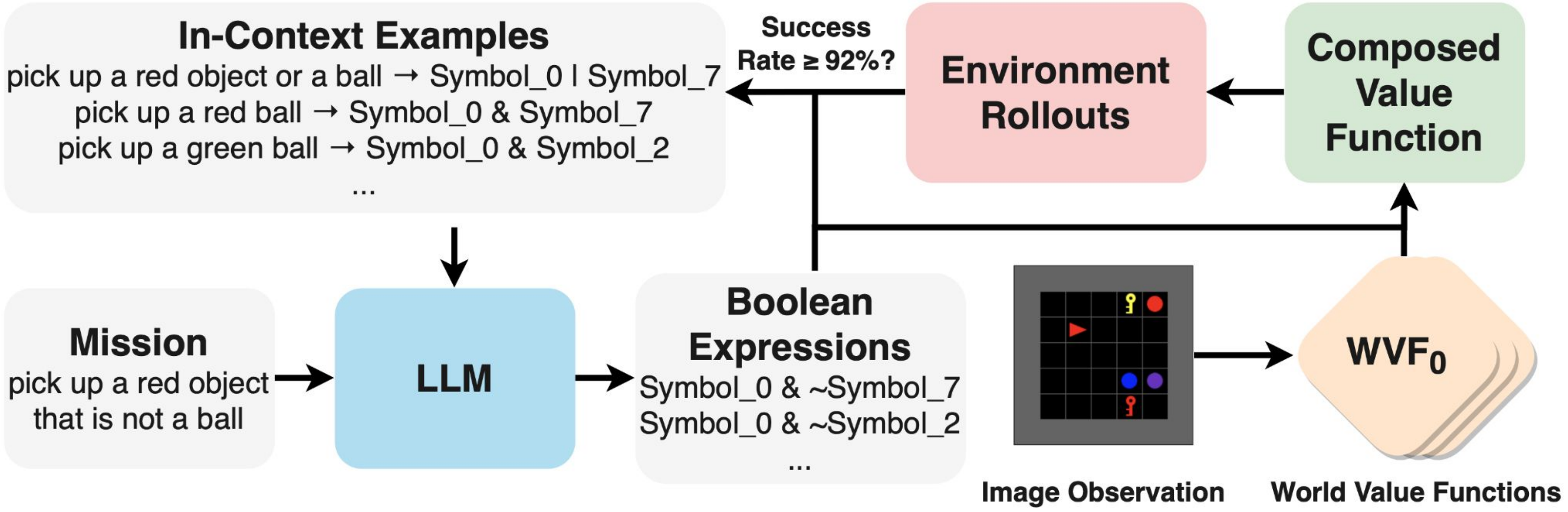
Compositionally-Enabled RL and Language Agent (CERLLA)



Compositionally-Enabled RL and Language Agent (CERLLA)



Compositionally-Enabled RL and Language Agent (CERLLA)



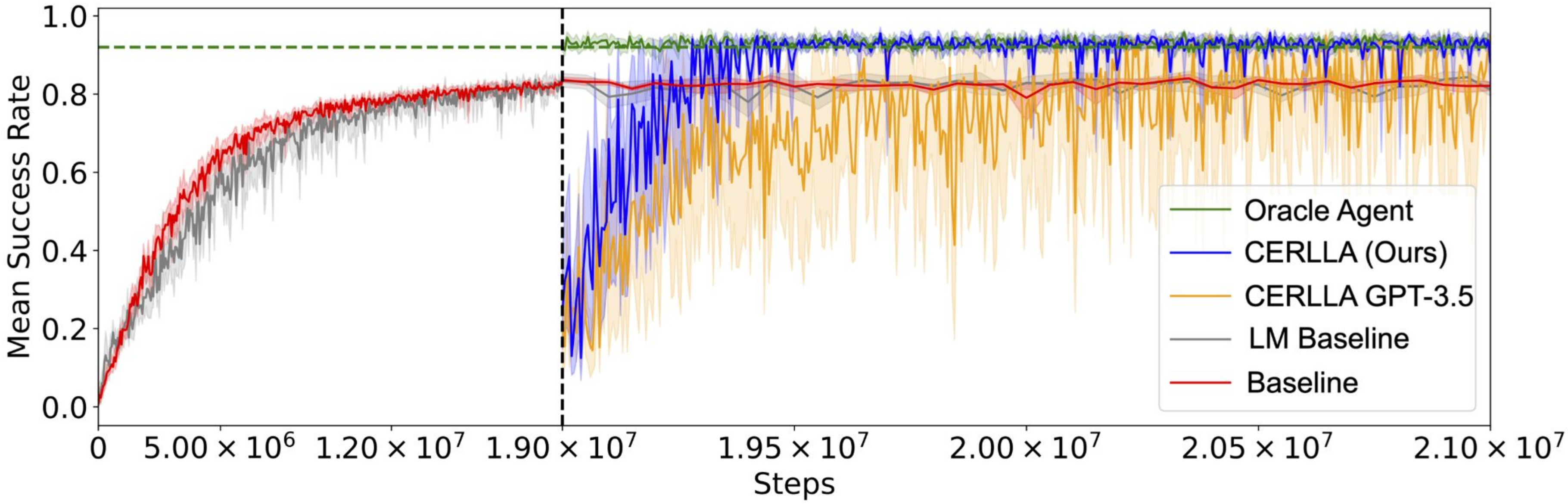
Experiments

- 162 tasks, learned simultaneously from vision and language.

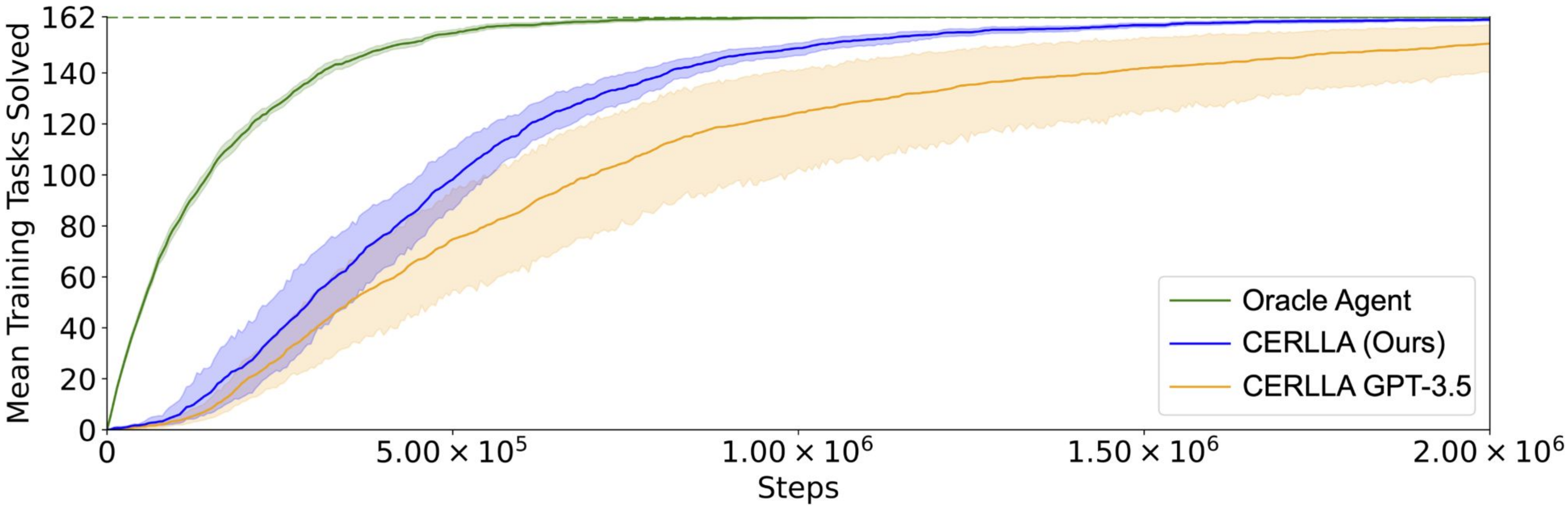
Experiments

- 162 tasks, learned simultaneously from vision and language.
- Evaluate **sample efficiency**, and **generalization**, comparing:
  - CERLLA (Ours): using OpenAI's GPT-4 LM
    - CERLLA GPT-3.5
  - Two non-compositional baseline DQNs
    - Baseline: RNN + CNN
    - LM Baseline: pretrained sentence embedding language model + CNN
  - Oracle Agent with access to the ground-truth compositional expressions for each task.

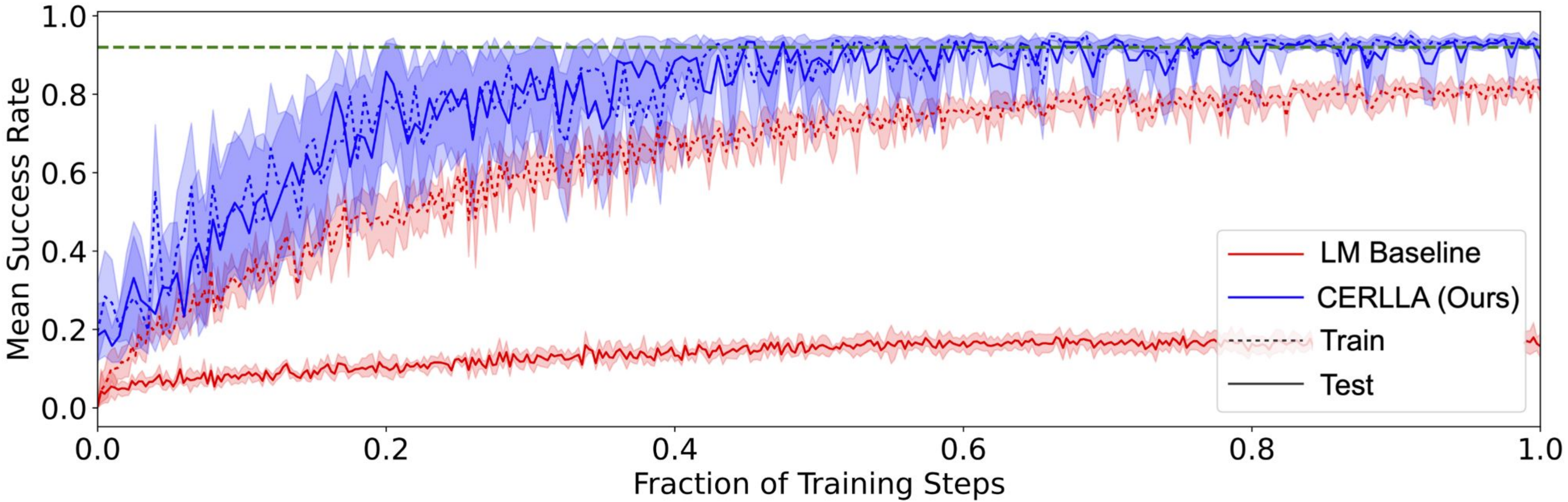
Sample Efficiency



Convergence



Generalization



Conclusion

- Introduces CERLLA, a **novel semantic parsing** method based on **in-context learning** and that learns from **environment feedback**.
- Simultaneously learns and solves a large collection of **162 compositional vision-language-RL tasks**.
- Outperforms non-compositional baselines with respect to sample efficiency and generalization to held-out tasks.

TMLR