Compositional Instruction Following with Language Models and Reinforcement Learning

Vanya Cohen[†1] Geraud Nangue Tasse[†2] Nakul Gopalan[3] Steven James[2] Matthew Gombolay[4] Raymond Mooney[1] Benjamin Rosman[2]

[1]The University of Texas at Austin [2]University of the Witwatersrand [3]Arizona State University [4]Georgia Institute of Technology

[†] Equal contribution

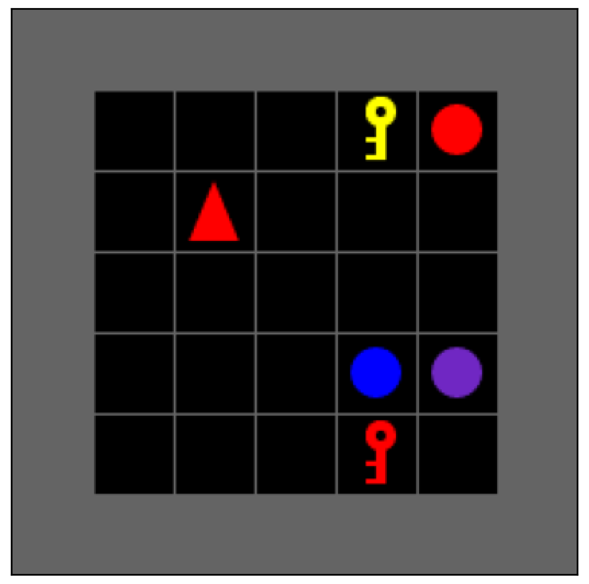
Motivation

- Combining reinforcement learning with language grounding is difficult because the agent must explore while mastering multiple language-conditioned tasks.
- We address this with the Compositionally-Enabled Reinforcement Learning Language Agent (CERLLA).
- CERLLA reduces sample complexity by leveraging compositional policy representations and a semantic parser trained via reinforcement learning and in-context learning.
- In a function-approximation setting, CERLLA exhibits compositional generalization to novel tasks.

Key Contributions

- CERLLA: a *compositionally-enabled* RL language agent with policies formed from conjunctions, disjunctions, and negations of pretrained compositional value functions.
- In-context learning + rollout feedback: improves the semantic parsing capabilities of an LLM.
- 162 unique tasks: solved in an augmented MiniGrid-BabyAI domain; to our knowledge, this represents the largest concurrently-learned compositional language-RL benchmark.

BabyAI Domain (Chevalier-Boisvert et al. (2019))



"Pick up the red key": the agent must combine red & key World Value Functions to solve the task.

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

World Value Functions (WVFs) are *goal-oriented* value functions that can be composed with logical operators such as $\wedge$, $\vee$, and $\neg$ to solve semantically meaningful tasks with no further learning. To achieve this, the reward function is extended to penalize the agent for attaining goals it did not intend:

$$\bar{r}(s,g,a) = \begin{cases} \bar{r}_{MIN} & \text{if } g \neq s \in \mathcal{G} \\ r(s,a) & \text{otherwise} \end{cases} \quad (1)$$

where $\bar{r}_{MIN}$ is a large negative penalty. The agent receives the unmodified reward $r(s,a)$ for all steps except where it reaches a different goal state than intended: $g \neq s \in \mathcal{G}$. Given $\bar{r}$, the related value function, termed a *world value function (WVF)*, can be written as

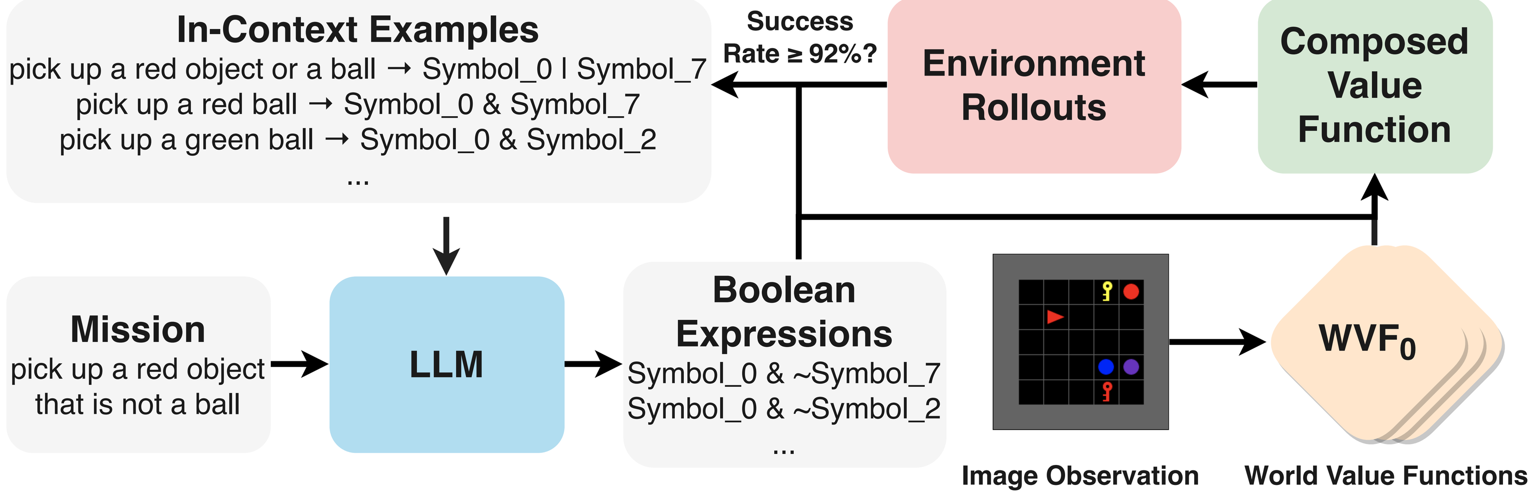
$$\bar{Q}(s,g,a) = \bar{r}(s,g,a) + \int_{\mathcal{S}} \bar{V}^{\bar{\pi}}(s',g)\, p(s' \mid s,a)\, ds' \quad (2)$$

Assume the agent has separately learned the task of collecting red objects (task R) and keys (task K). Using these value functions, the agent can immediately solve the tasks defined by their union ($R \vee K$), intersection ($R \wedge K$), and negation ($\neg R$) as follows:

$$\bar{Q}^*_{R\vee K} = \bar{Q}^*_R \vee \bar{Q}^*_K \coloneqq \max\{\bar{Q}^*_R, \bar{Q}^*_K\},$$
$$\bar{Q}^*_{R\wedge K} = \bar{Q}^*_R \wedge \bar{Q}^*_K \coloneqq \min\{\bar{Q}^*_R, \bar{Q}^*_K\},$$
$$\bar{Q}^*_{\neg R} = \neg\bar{Q}^*_R \coloneqq \left(\bar{Q}^*_{MAX} + \bar{Q}^*_{MIN}\right) - \bar{Q}^*_R,$$

where $\bar{Q}^*_{MAX}$ and $\bar{Q}^*_{MIN}$ are the WVFs for the *maximum* and *minimum* tasks, respectively.

Compositionally-Enabled Reinforcement Learning Language Agent (CERLLA)



Pipeline overview: Agent receives a BabyAI command + 10 BM25-retrieved examples, generates 10 Boolean parses, tests each for 100 roll-outs, and retains those with $\geq 92\%$ as new in-context examples.

Example language instructions and corresponding Boolean expressions for the *yellow* and *box* attributes.

Language Instruction	Ground Truth Boolean Expression
pick up a yellow box	$yellow\ \&\ box$
pick up a box that is not yellow	$\sim yellow\ \&\ box$
pick up a yellow object that is not a box	$yellow\ \&\ \sim box$
pick up an object that is not yellow and not a box	$\sim yellow\ \&\ \sim box$
pick up a box or a yellow object	$yellow \mid box$
pick up a box or an object that is not yellow	$\sim yellow \mid box$
pick up a yellow object or not a box	$yellow \mid \sim box$
pick up an object that is not yellow or not a box	$\sim yellow \mid \sim box$
pick up a box	box
pick up an object that is not a box	$\sim box$
pick up a yellow object	$yellow$
pick up an object that is not yellow	$\sim yellow$

The prompting strategy for the CERLLA semantic parsing module.

Role	Content
System	"We are going to map sentences to Boolean expressions. The Boolean expression variable Symbols are numbered 0 to 8, e.g. $Symbol\_0$, $Symbol\_1$... The operators are and : &, or : |, not : ~. I will now give a new sentence and you will come up with an expression. Now wait for a new sentence command. Respond with a list of 10 candidate Boolean expressions. Respond only with the list of Boolean expressions. Never say anything else."
User (Example) *Assistant*	"pick up a red ball" "$Symbol\_0\ \& Symbol\_7$"
	[Additional in-context examples]
User (Command) *Assistant*	"pick up a red object that is not a ball" "$Symbol\_0\ \&\ Symbol\_1\ \&\ \sim Symbol\_2$" "$Symbol\_3\ \&\ \sim Symbol\_4$" "$Symbol\_5\ \&\ Symbol\_6\ \&\ \sim Symbol\_7$" [Additional candidate expressions]

References

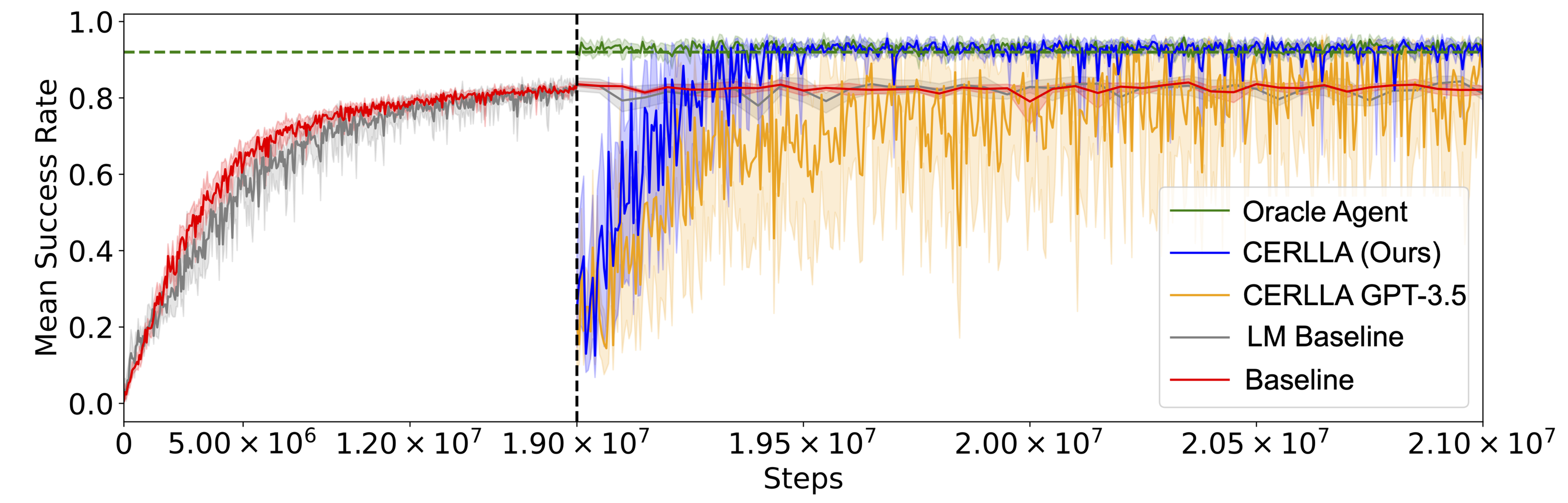
Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: First steps towards grounded language learning with a human in the loop. In *International Conference on Learning Representations*, 2019.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. World value functions: Knowledge representation for multitask reinforcement learning. In *The 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2022.

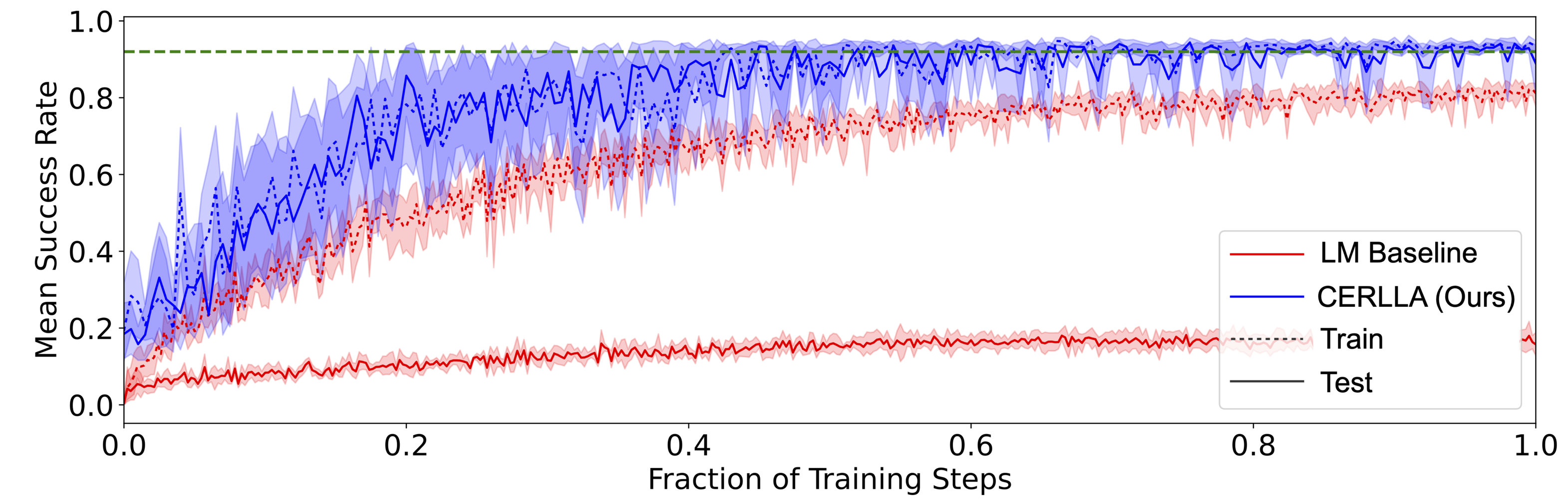
Key Findings

- Compositionality improves performance and sample efficiency even when accounting for the pretraining steps of the World Value Functions.
- CERLLA generalizes systematically to held-out tasks by leveraging compositional structure.
- CERLLA converges to the performance of an Oracle Agent which has access to the correct compositions of the World Value Functions to complete each task.

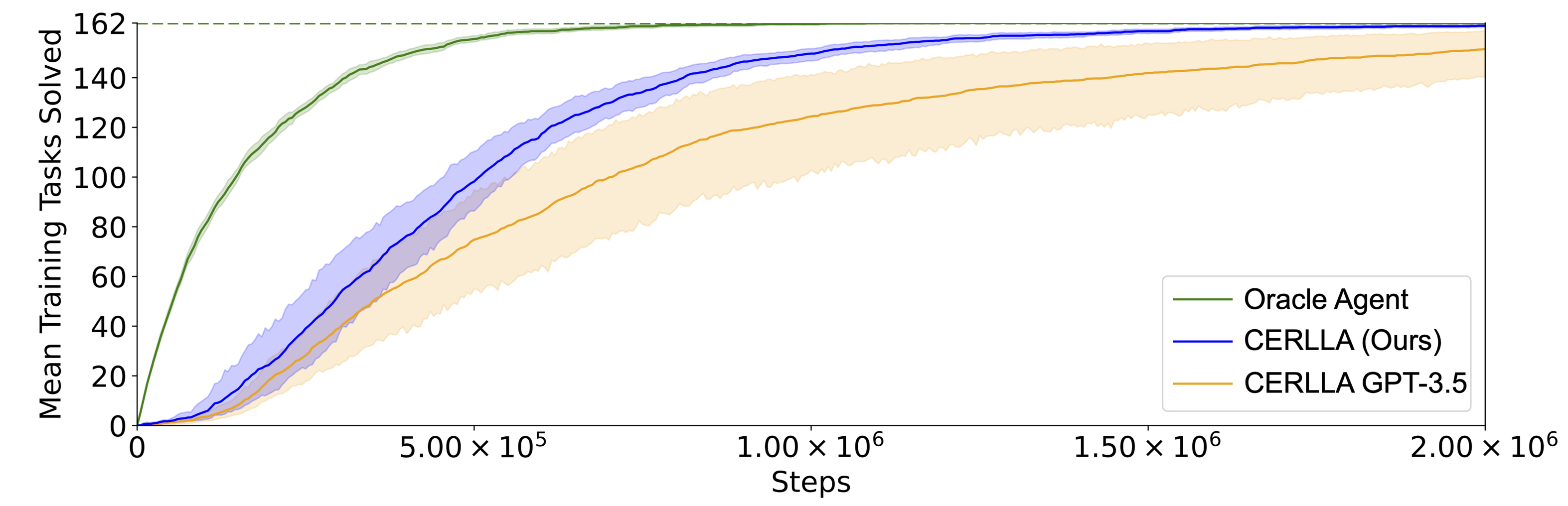
Results



162-task learning: CERLLA reaches 92% success in 0.6M env-steps (19 M counted pre-train); baseline stalls near 80 %. Dashed line shows the Oracle upper bound.



Held-out generalization: CERLLA generalizes from 81 train tasks to 81 held-out test tasks, while the non-compositional LM baseline exhibits limited generalization.



Solved-task count: CERLLA rapidly acquires all tasks in the environment; logistic shape reflects a shrinking unsolved pool.

vanya@utexas.edu