Empowering LLM Agents with Zero-Shot Optimal Decision-Making through Q-learning

Jiajun Chai, Sicheng Li, Yuqian Fu, Yuanheng Zhu, Dongbin Zhao

Institution of Automation, Chinese Academy of Sciences

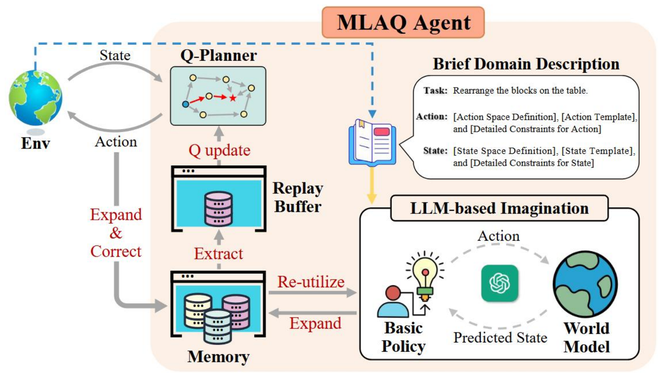
School of Artificial Intelligence, University of Chinese Academy of Sciences

{chaijiajun2020, lisicheng2024, fuyuqian2022, yuanheng.zhu, dongbin.zhao}@ia.ac.cn

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- **Generalization.** LLM agents leverage the general ability of LLMs to make **zero- or few-shot** decisions but fail in making optimal decisions.
- **Optimization.** Agents trained via reinforcement learning (RL) could make optimal decisions but require extensive environmental data.
- **Overall.** We combine the zero-shot of LLMs with the optimality of RL. We construct an imagination space fully based on LLM to perform imaginary interactions for deriving zero-shot policies, and employ Q-learning to derive optimal policies from transitions within memory.

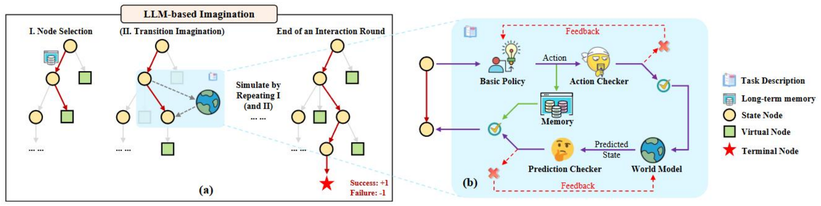


The model-based framework for MLAQ

Contribution

We propose **Model-based LLM Agent with Q-Learning (MLAQ)**, a novel LLM agent framework without accessing any environmental tools. The main contribution is three-fold.

- **Zero- or Few-Shot Optimization with LLMs.** MLAQ integrates a Q-planner, memory, and imagination space to implement decision-making and optimization paradigms with MDP properties in natural language.
- **Efficient Exploration in Imagination Space.** An MCTS-style planning approach is proposed to balance exploration and exploitation within the imagination space. We efficiently guide the exploration without the need for environmental tools, **ensuring a sub-linear regret bound**.
- **Against Hallucination.** A mixed-examine mechanism is proposed to improve the quality of imaginary transitions. It uses LLM-based self-examine to eliminate erroneous transitions from the memory and refines the world model.



The planning approach in the imagination space. (a) Imaginary interaction process. (b) Imagine a new transition with self-examine.

Method

Optimizing LLM Agent with Q-Planner

We develop an RL-style LLM agent framework, which contains a memory module, a Q-Planner, and an LLM-based imagination space.

- **Q-Planner.** Applying Q-learning to a task-specific replay buffer D extracted from the domain memory, aiming to reduce the exploration space.
- **Replay Buffer.** An initial replay buffer stores the transitions along the optimal trajectory, and the agent keeps exploring to expand it.
- **Imaginary Interaction.** An LLM-based basic policy and world model are established to perform imaginary interactions.
- **Environmental Interaction.** MLAQ agent explores the imagination space to derive the optimal policy using the Q-planner, and then outputs actions to interact with the environment.

LLM-based Imagination for MLAQ

- **Exploration in LLM-based Imagination Space.** We present an MCTS-style planning method that balances exploration and exploitation while using only LLMs to generate imaginary transitions.
- In short, we propose a variant of UCB specifically designed for exploration in LLM agents with **unknown available actions**, and theoretically satisfies a sub-linear regret bound:

$$a^* = \operatorname{argmax}_{a\in\hat{\delta}(s)} \mathrm{vUCB}(s,a) = \operatorname{argmax}_{a\in\hat{\delta}(s)} \left[V\big(c(s,a)\big) + f\left(N(s), N\big(c(s,a)\big)\right)\right]$$

- **Overall.** Through the interaction between LLM based basic policy and world model, a large number of imaginative transitions were generated without contact with the environment, and mixed examination was used to improve the accuracy of these transitions as much as possible. Q-learning was then used to obtain the optimal strategy.

Experiments

- **Environments.** We conduct experiments on the BlocksWorld benchmark for single-agent setting and the RoCo-benchmark for multi-agent setting.
- **Baselines.** We compare MLAQ with CoT, RAP, Rex, RAFA, and RoCo. In line with RAP, we group all tasks by their optimal steps, indicating the length of the tasks' optimal decision sequences.
- **Experimental Results.** MLAQ truly achieves zero- or few-shot optimal decision-making without using environmental tools, and its performance far exceeds existing LLM agents

Methods	2-step	4-step	6-step	8-step	10-step	12-step
CoT	0.22	0.14	0.02	0.02	0.00	0.00
REX	0.80	0.45	0.25	-	-	-
RAFA	-	0.97	0.75	-	-	-
RAP	0.67	0.76	0.74	0.48	0.17	0.09
MLAQ	**1.00**	**1.00**	**1.00**	**0.97**	**0.93**	**0.90**

Experiment results on BlocksWorld

Metrics	Methods	1-step	2-step	3-step	4-step	5-step	6-step	Average
Success Rate	RoCo	1.00	0.64	0.47	0.10	0.03	0.00	0.35
	MLAQ	**1.00**	**0.96**	**0.97**	**1.00**	**0.93**	**1.00**	**0.98**
Env Replans (n-shot)	RoCo	0.30	6.30	5.60	9.74	7.67	6.92	6.41
	MLAQ	**0.00**	**0.04**	**0.03**	**0.10**	**0.10**	**0.04**	**0.06**
Optimal Rate	RoCo	0.80	0.36	0.27	0.00	0.00	0.00	0.21
	MLAQ$^-$	0.95	0.64	0.57	0.67	0.33	0.43	0.58
	MLAQ	**1.00**	**0.86**	**0.77**	**0.73**	**0.50**	**0.75**	**0.75**
Average Token	RoCo	10605	530817	332730	305175	345762	320045	322630
	MLAQ$^-$	**7093**	**15104**	**16436**	**22197**	**18133**	**8220**	**15216**
	MLAQ	8491	66367	156490	119916	409495	243151	175560
Optimal Gap	RoCo	0.35	3.37	3.43	3.33	2.97	2.00	2.72
	MLAQ$^-$	0.10	0.64	0.73	0.53	1.20	0.46	0.65
	MLAQ	**0.00**	**0.32**	**0.40**	**0.43**	**0.80**	**0.25**	**0.39**
Memory Re-Util. Ratio		0.27	0.61	0.71	0.72	0.73	0.83	0.66

Experiment results on RoCo-Benckmark (Sort domain)

Ablation Methods	Optimal Rate	Token	Env Replans
RAP	0.48	-	-
MLAQ w/o domain memory	0.87	378941	0.02
MLAQ w/o self-examine	0.57	20115	0.81
MLAQ w/o env-examine	0.95	40772	0.05
MLAQ	1.00	39688	0.02

Ablation results about domain memory, self-examination, and env-examination

Interactive Website: http://mlaq.site

Code Link: https://github.com/laq2024/MLAQ