Divide and Conquer: Grounding LLMs as Efficient Decision-Making Agents via Offline Hierarchical Reinforcement Learning

Zican Hu[12], Wei Liu[3], Xiaoye Qu[2], Xiangyu Yue[4], Chunlin Chen[1], Zhi Wang[12] ✉, Yu Cheng[4] ✉

[1] Nanjing University [2] Shanghai AI Laboratory [3] HKUST [4] CUHK ✉ Corresponding Authors

Contact & Cooperation

zhiwang@nju.edu.cn

Challenge & Motivation

➢ **Deficient exploration**:

Difficulty in effectively exploring vast action spaces for long-horizon tasks.

➢ **Inadequate credit assignment** :

Inadequate credit assignment in sparse-reward scenarios.

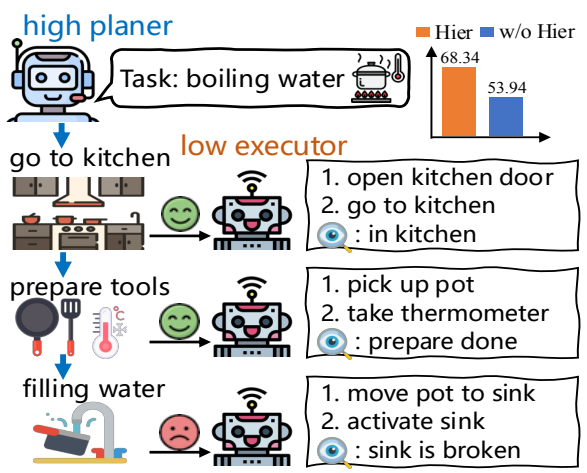
Solution & Contribution

➢ **Hierarchical framework**:

We propose a hierarchical framework with with superior parameter efficiency to solve complex decision tasks.

➢ **Offline-to-online adaptation**:

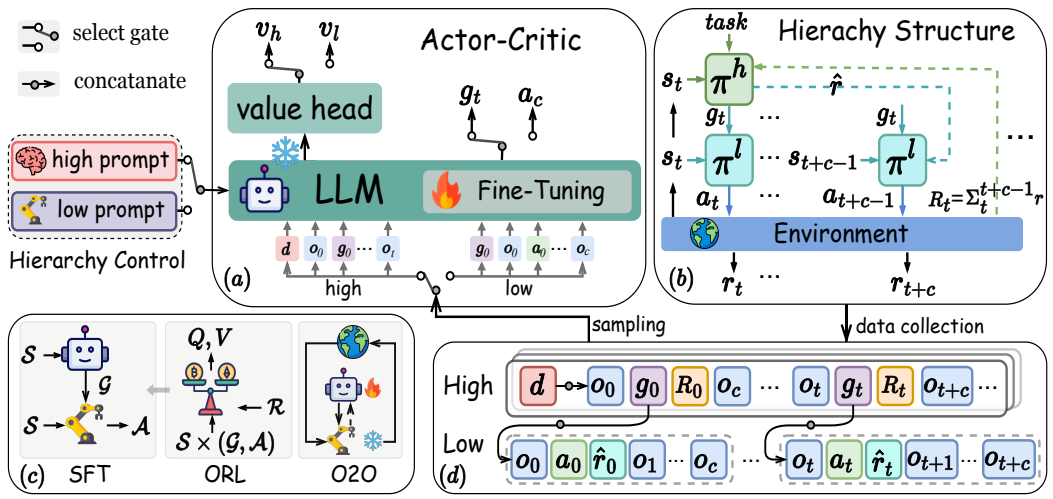
Our approach enables rapid adaptation to non-stationary environments through generalizable hierarchical skills.



Insights illustrated via a "boiling water" case

Method & Framework

➢ Grounding Language Models as Efficient Decision-Making Agents via Offline HiErarchical RL Framework (GLIDER)



a) Hierarchical Actor-Critic architecture with prompt-controlled high- and low-level training.

- *Sentence-Level Critic :*

$$\mathcal{L}_Q(\phi) = \mathbb{E}_{(s,u,r,s')\sim D_r}[(r + \gamma V_{\bar{\psi}}(s') - Q_\phi(s,u))^2]$$
$$\mathcal{L}_V(\psi) = \mathbb{E}_{s\sim D_r}[\mathbb{E}_{u\sim\pi_\theta(\cdot|s)}[L_2^\tau(Q_{\bar{\phi}}(s,u) - V_\psi(s))]]$$

- *Token-Level Actor :*

$$\mathcal{L}_\pi(\theta) = -\mathbb{E}_{(s,u)\sim D_r}[\exp(\frac{1}{\lambda}A(s,u)) \cdot \sum_{i=1}^{n} log\ \pi_\theta(w_i|s, w_{1:i-1})]$$

b) Hierarchical policy where the high-level π^h generates sub-task g only when the low-level π^l executes primitive actions for c steps. c could differ across sub-tasks.

c) The training pipeline comprises SFT, ORL (offline RL), and O2O (offline-to-online RL) stages.

d) Hierarchical trajectories composed as:

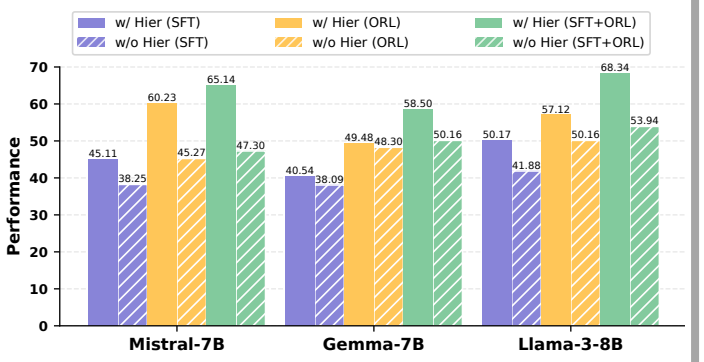
$$D^h = \sum_N [d; (o_0, g_0, \sum r_{0:c-1}, o_c), \dots, (o_t, g_t, \sum r_{t:t+c-1}, o_{t+c}), \dots]$$
$$D^l = \sum_N \sum_t [g_t; (o_t, a_t, \hat{r}_t, o_{t+1}), \dots, (o_{t+c-1}, a_{t+c-1}, \hat{r}_{t+c-1}, o_{t+c})]$$

Experiment & Analyzation

➢ Result across three backbone on two decision benchmarks.

Backbone	Method	ScienceWorld Seen	ScienceWorld Unseen	AlfWorld Seen	AlfWorld Unseen
Mistral-7B	ReAct	20.72	17.65	7.86	5.22
	Reflexion	21.07	18.11	11.56	6.00
	SwitchSage	48.40	45.25	30.29	26.52
	NAT	57.12	50.79	64.43	68.96
	ETO	58.17	51.85	66.84	71.43
	GLIDER	**67.31** (↑ 15.71%)	**65.14** (↑ 25.63%)	**70.02** (↑ 4.76%)	**74.83** (↑ 4.76%)
Gemma-7B	ReAct	3.58	3.51	6.43	2.24
	Reflexion	4.94	3.93	7.14	2.99
	SwitchSage	33.43	30.90	8.23	5.72
	NAT	47.63	44.98	67.86	65.88
	ETO	50.44	47.84	66.43	68.66
	GLIDER	**63.67** (↑ 26.23%)	**58.50** (↑ 22.28%)	**72.12** (↑ 6.28%)	**70.88** (↑ 3.23%)
Llama-3-8B	ReAct	24.76	22.66	2.86	3.73
	Reflexion	27.23	25.41	4.29	4.48
	SwitchSage	42.22	40.58	20.39	10.78
	NAT	55.24	48.76	60.71	59.70
	ETO	57.90	52.33	64.29	64.18
	GLIDER	**77.43** (↑ 33.73%)	**68.34** (↑ 30.59%)	**71.56** (↑ 11.31%)	**75.38** (↑ 17.45%)

➢ Ablation performance across model architectures and pipeline.



➢ Online fine-tuning result against AC and AWAC baselines.