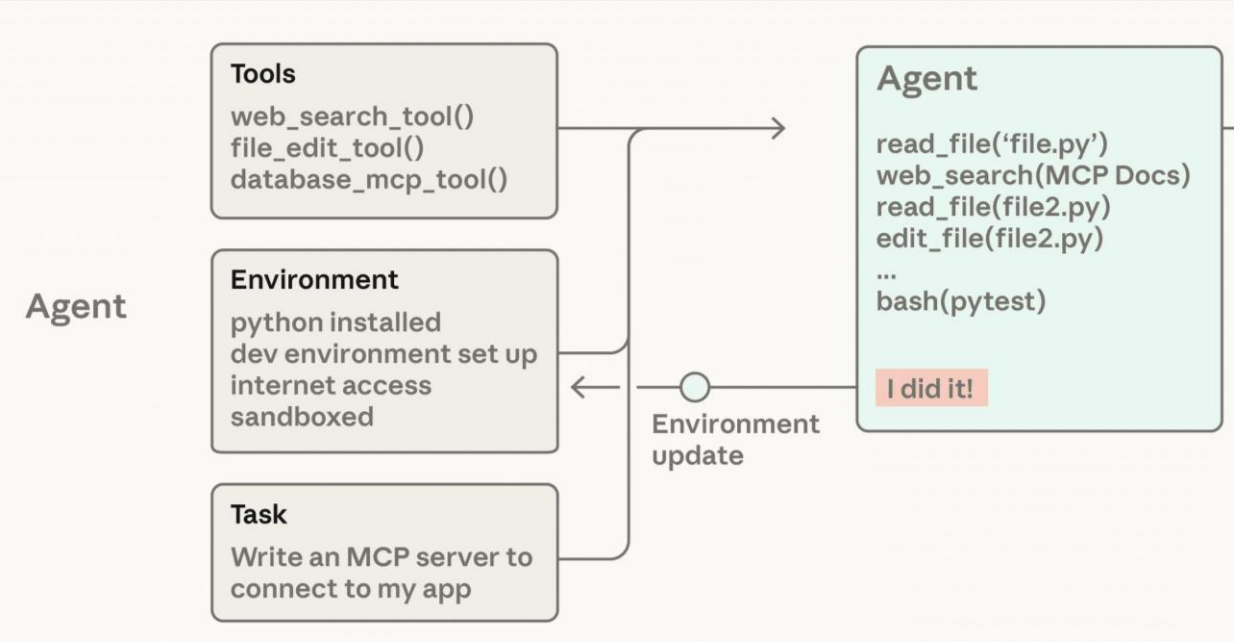
AgentGym-RL: An Open-Source Framework to Train LLM Agents for Long-Horizon Decision Making via Multi-Turn RL

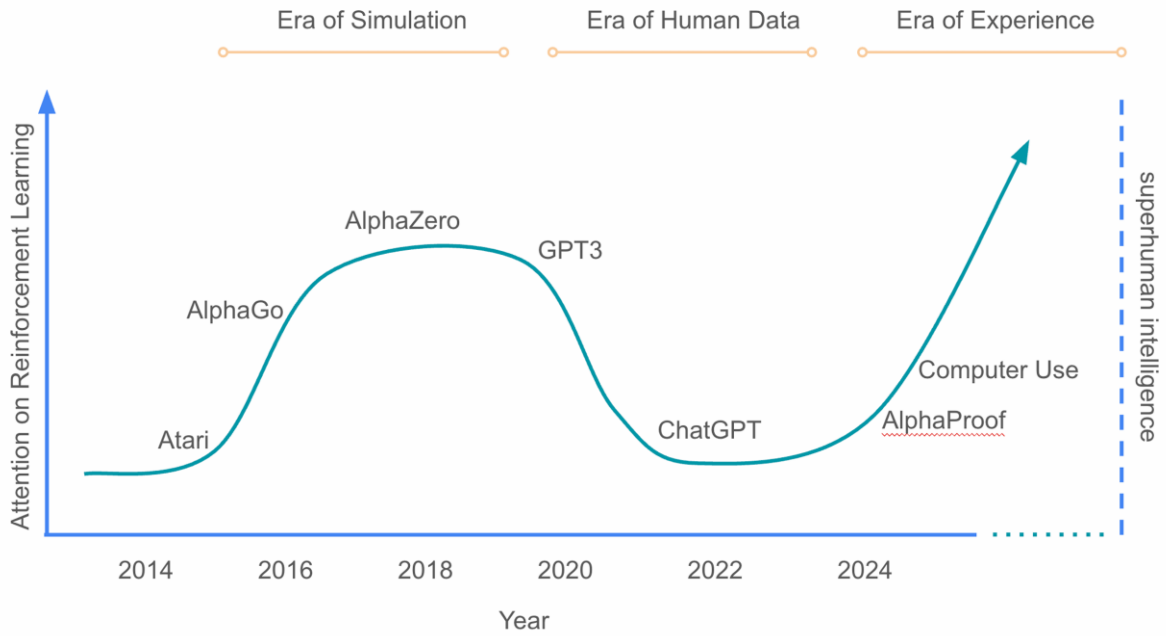
Zhiheng Xi

Fudan University

Email: zhxi22@m.fudan.edu.cn

Motivation & Background



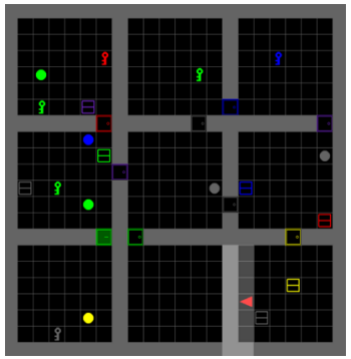


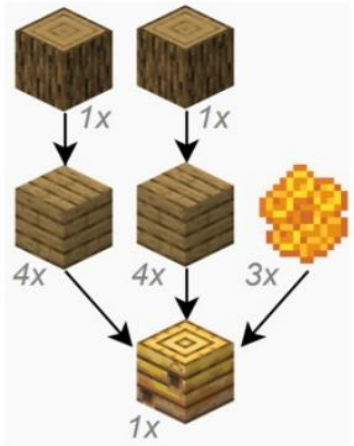
- Agents must acquire skills by **actively exploring** environment
- **Reinforcement Learning** is the natural optimizer

Motivation & Background

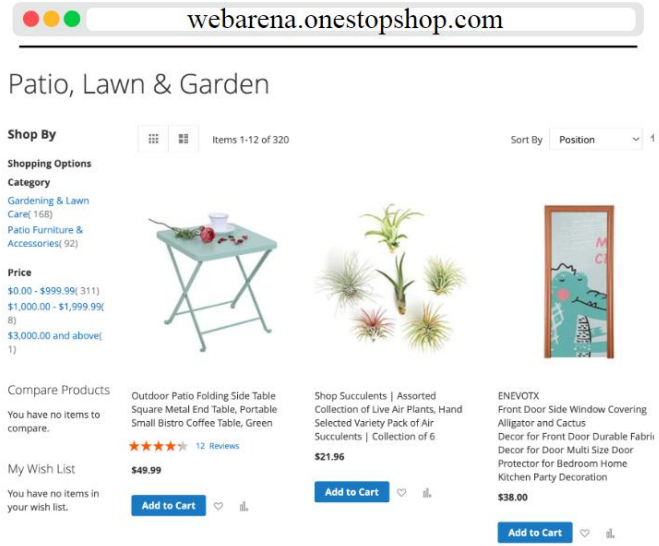
- Existing works: Restricted complexity, diversity

Toy Environment Setting





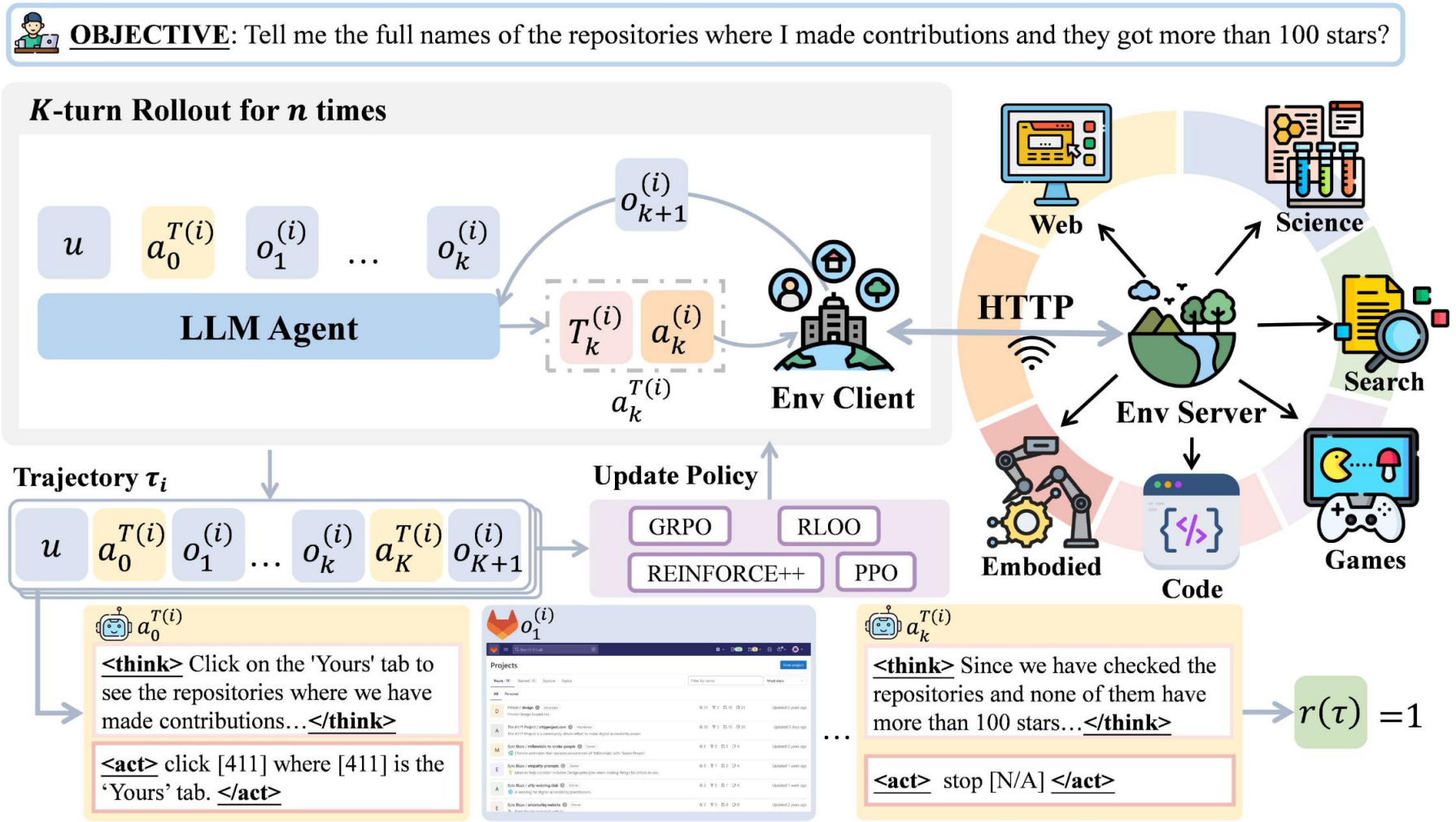
Real-World Scenario



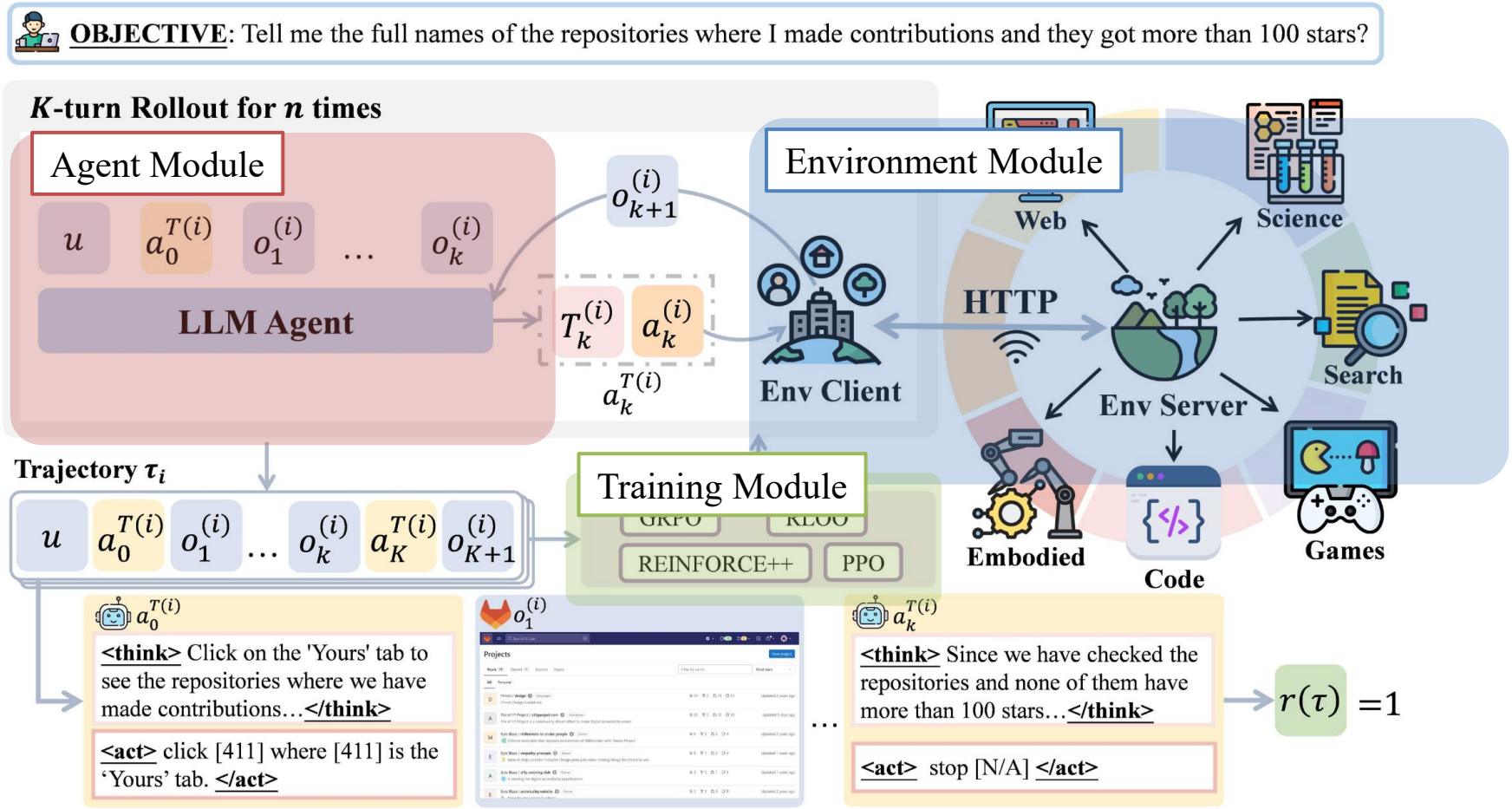
- Agents that perform well only in toy settings **struggle to transfer to real-world** scenarios
- **Diversity** in environments is a prerequisite for their **generalization**

Framework: AgentGym-RL

Overview

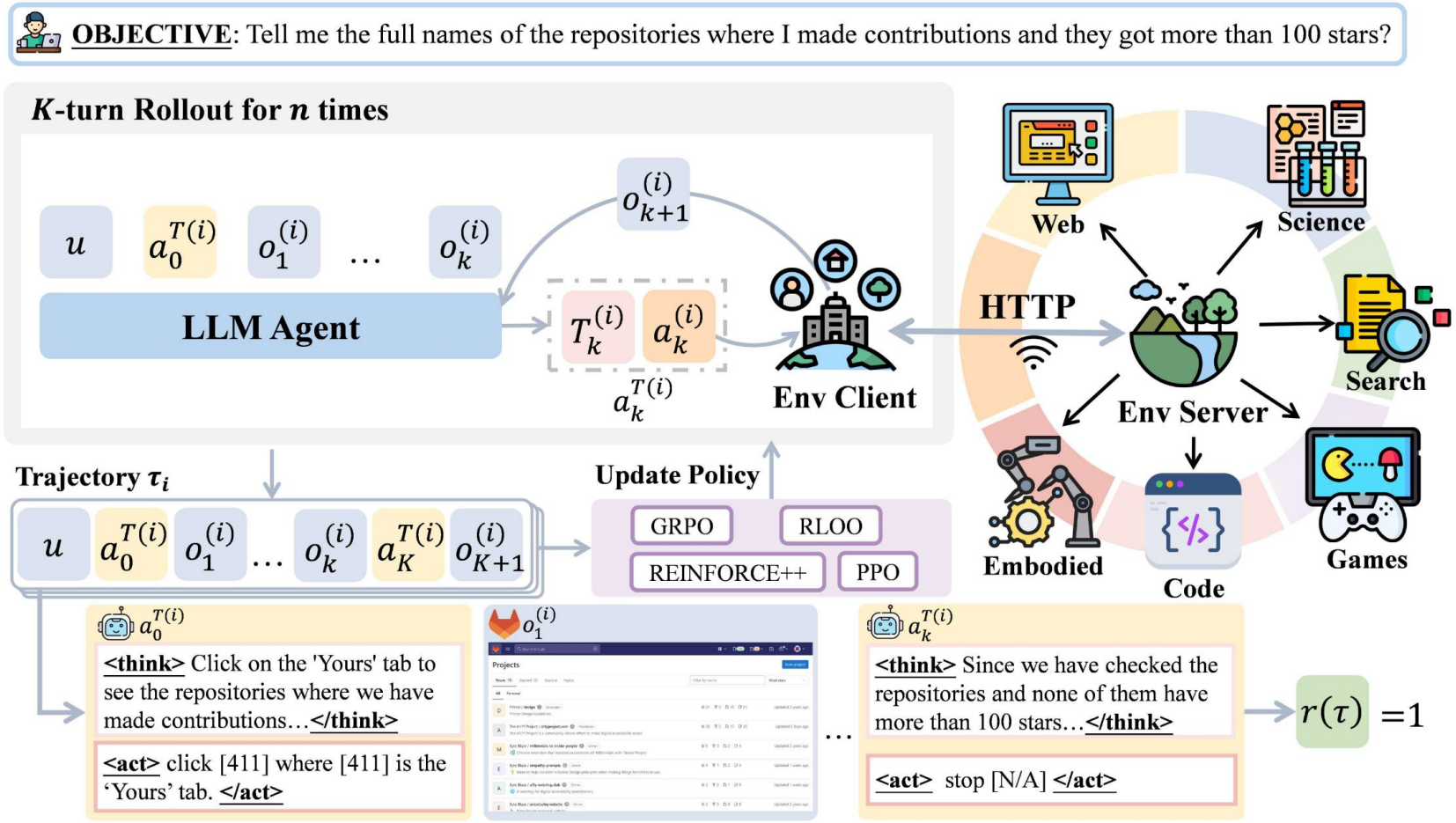


Framework: AgentGym-RL



- **Feature: Modular Design**

Framework: AgentGym-RL

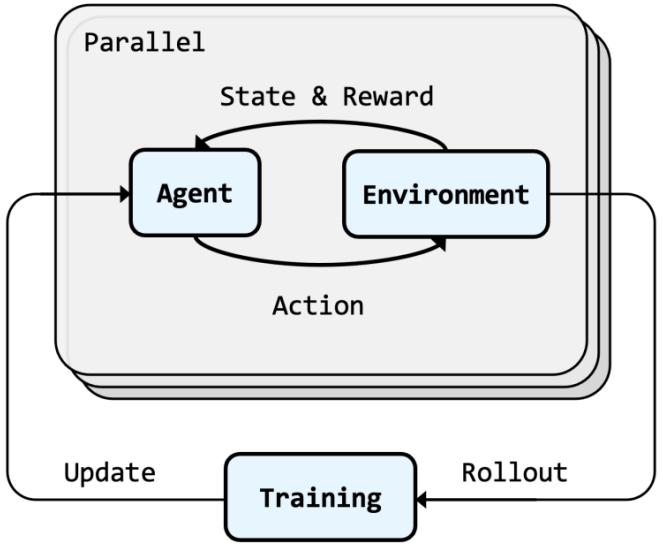


- **Feature: Comprehensive algorithm support.**
  - **RL algorithms:** GRPO, PPO…
  - **Offline paradigms:** SFT, DPO…
- **Feature: Diverse scenarios and environments.**
  - BabyAI, TextCraft, SciWorld, SearchQA, WebArena

Framework: AgentGym-RL

```
# Stage 1: Generate responses
task_ids = expand(task_ids, sample_num)
envs = create_env_clients(task_ids, “webarena”, base_url)

Do in parallel:
    for (env, task_id) in zip(envs, task_ids):
        env.reset(task_id)

handlers =[
    RolloutHandler().add_user_message(env.observe())
    for env in envs]

for i in range(max_rounds)
    prompts = [h.get_prompt() for h in handlers]
    responses = actor.generate(prompts)

    results = thread_safe_list()
    Do in parallel:
        for (env, response) in zip (envs, responses):
            results.append(env.step(response))

    for (h, r, res) in zip(handlers, responses, results):
        h.add_assistant_message(r)
        h.add_user_message(res.state)
        h.score = res.score

    if all_done(handlers): break
```

```
# Stage 2: Prepare experience
batch = gen_batch_from_rollout_handlers(handlers)
batch = actor.compute_log_prob(batch)
batch = reference.compute_ref_log_prob(batch)
batch = compute_advantages(batch, method="grpo")

# Stage 3: Actor training
actor.update_actor(batch)
```

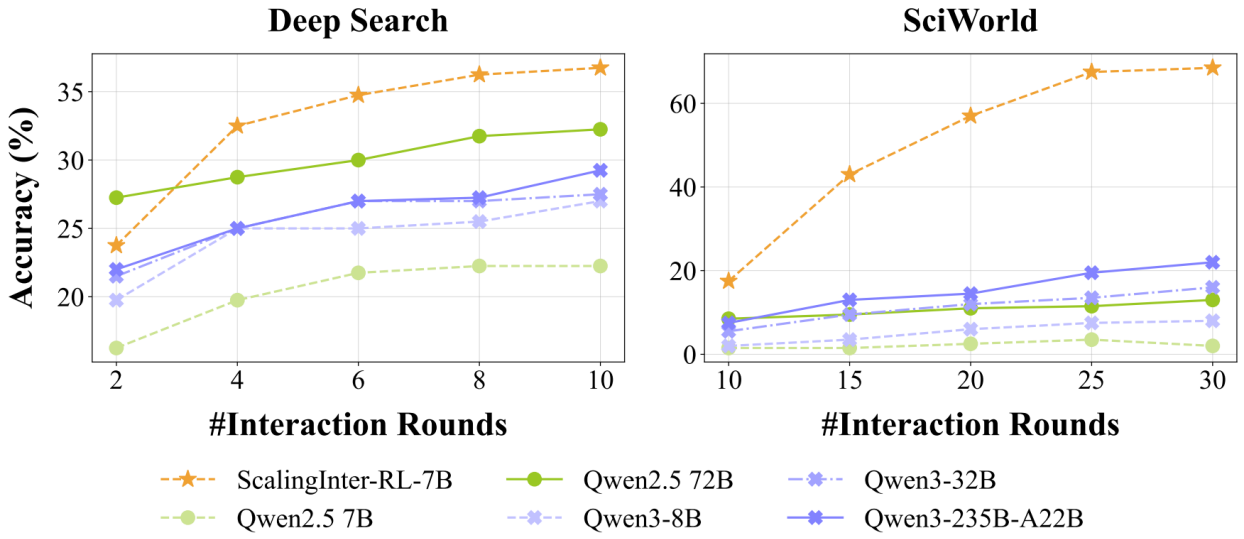


- **Engineering optimizations**
  - **Extensibility**: plug-and-play design **Scalability**: parallel rollout **Reliability**: address memory leaks

Method: ScalingInter-RL

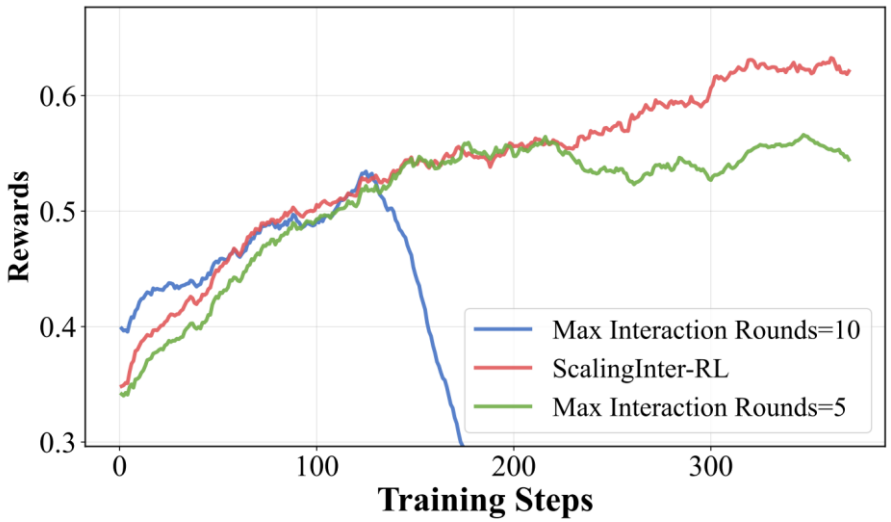
Insight: effective progress requires **expanding external interactions** with the environment

- **Experiment**: Scaling test-time interaction turns



- Long-horizon interaction **enhance performance**
- **Plateau** as the number of interactions continues to grow

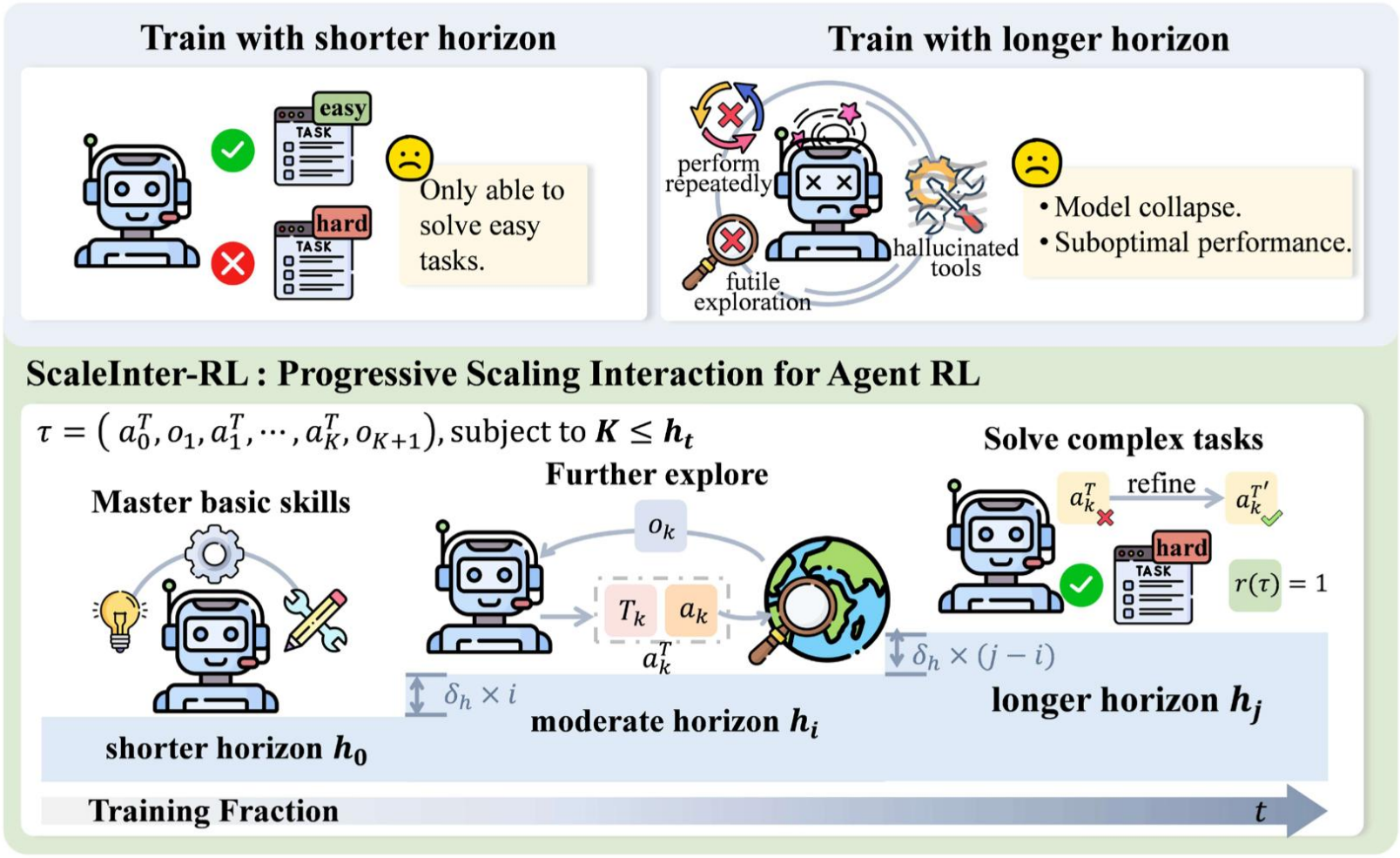
- **Experiment**: Vanilla long-horizon RL



- Larger interaction horizons introduce **training instability**, leading to training collapse

Method: ScalingInter-RL

- **ScalingInter-RL**: **gradually** increases the number of interaction turns during RL training

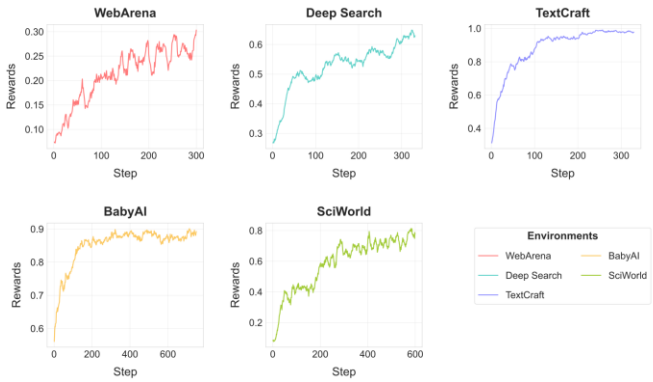


- Horizon is updated every Δ steps:

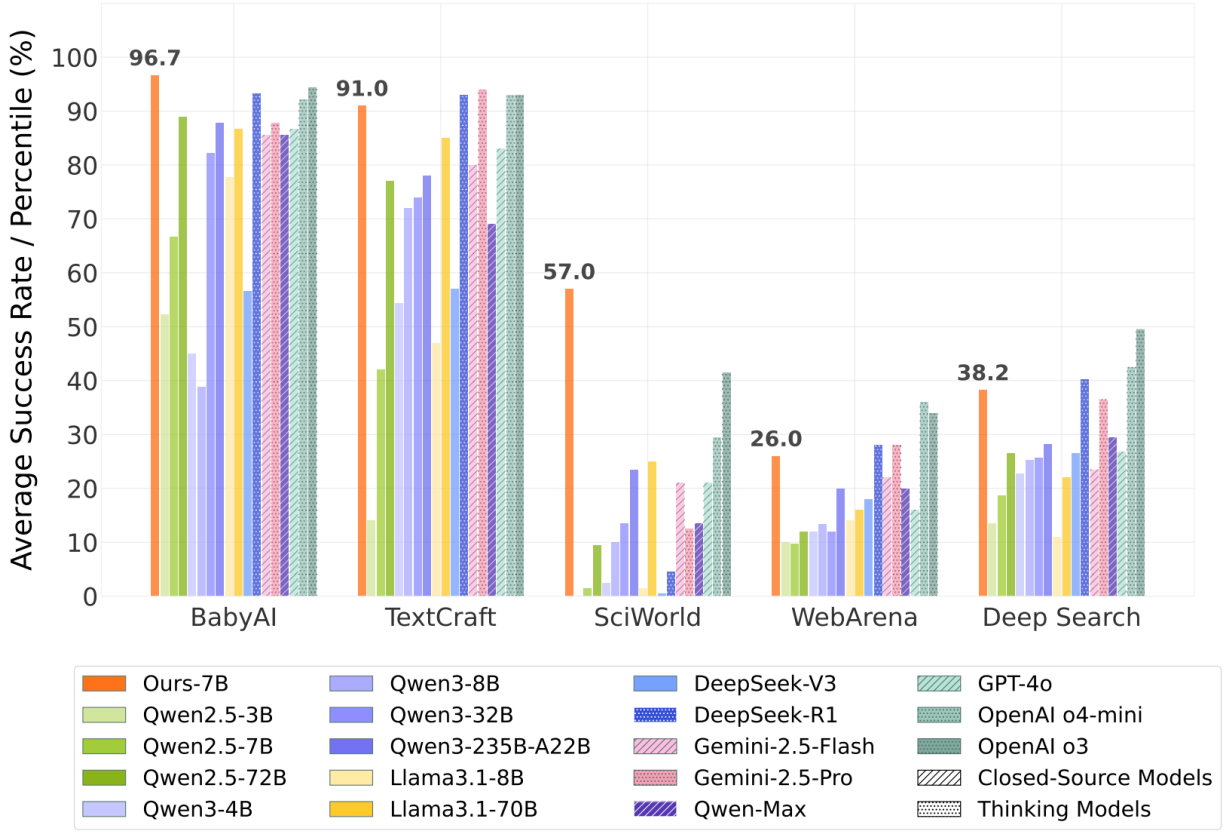
$$h_{t+1} = h_t + \delta_h$$

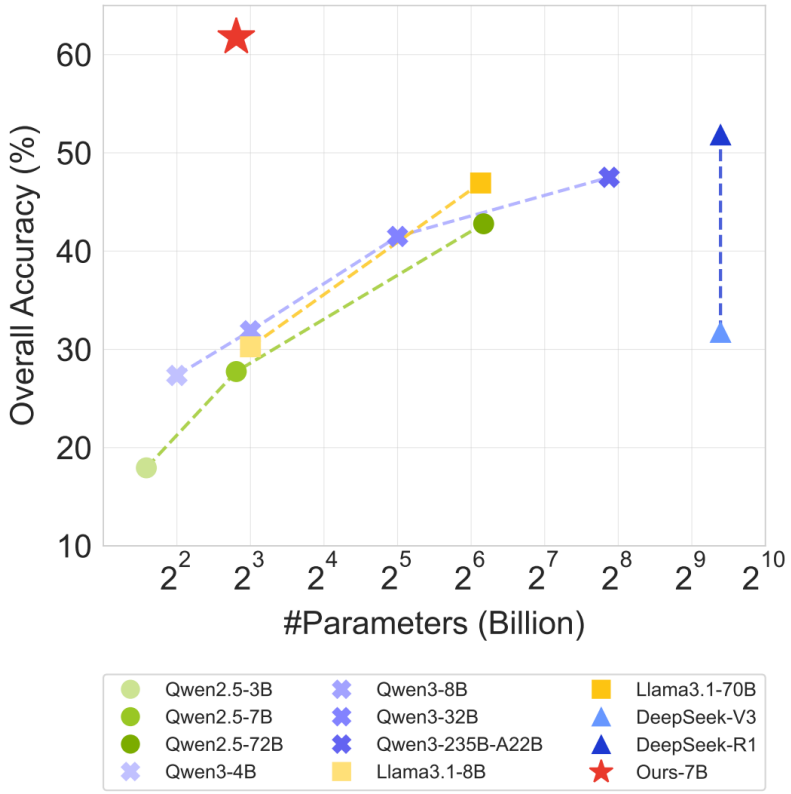
  - Begin with a short interaction horizon: focuses on **exploitation**
  - Horizon expands: encourage to **explore**

- Stable, consistent optimization



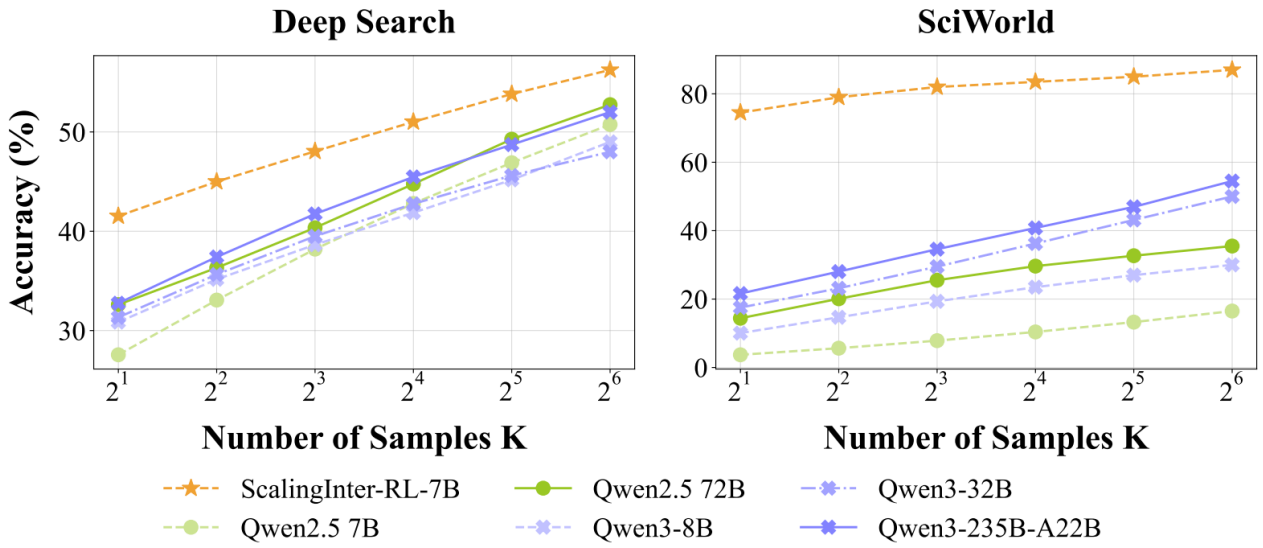
Experiment Results





- Our RL model **outperforms other open-source** models by a large margin
- Leads **in average success rate over closed-source** models like GPT-4o and Gemini-2.5-Pro

Experiment Results



- Pass@K performance

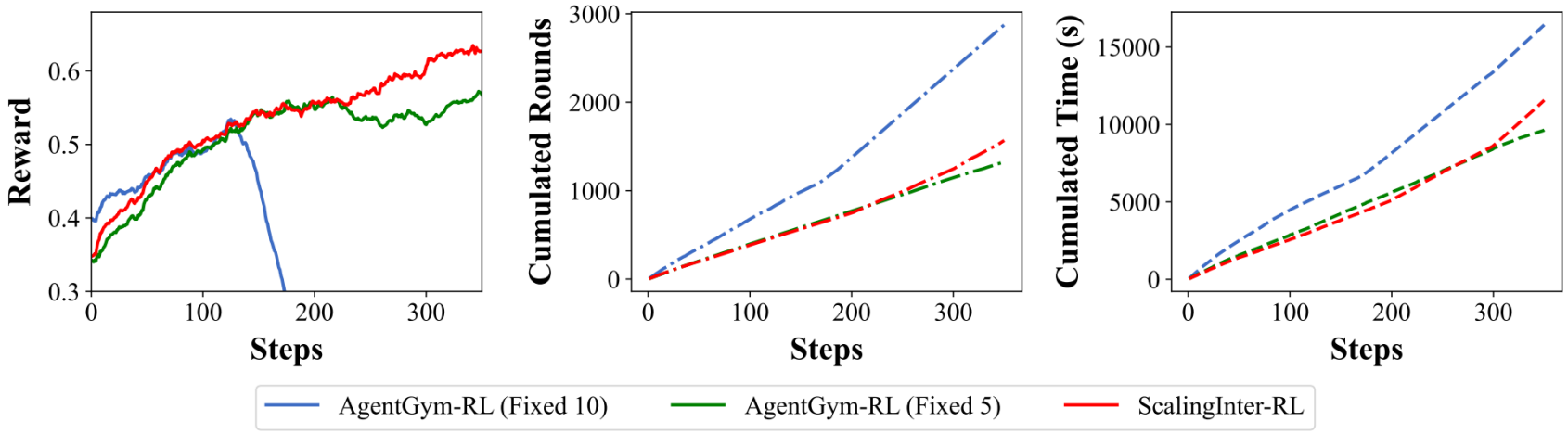
RL Algorithms	TextCraft	BabyAI	SearchQA
Qwen2.5-3B-Instruct			
GRPO	**75.00**	**93.33**	**25.75**
REINFORCE++	28.00	70.00	13.25
Qwen2.5-7B-Instruct			
GRPO	**89.00**	**92.22**	**34.00**
REINFORCE++	73.00	84.44	24.00

- Evaluation results of different RL algorithms.

Experiment Results

RL Algorithm	Method	TextCraft	BabyAI	SciWorld
Base Model	-	42.00	66.67	1.50
PPO	AgentGym-RL-7B	68.00	86.66	10.83
	ScalingInter-7B	**71.00**	**90.00**	**25.69**
REINFORCE++	AgentGym-RL-7B	73.00	84.44	13.63
	ScalingInter-7B	**77.00**	**87.77**	**26.14**

- Applying ScalingInter-RL to more algorithms.



- Efficiency analysis of ScalingInter-RL

Conclusion

- **AgentGym-RL**: a unified reinforcement learning framework for training LLM agents in long-horizon, multi-turn decision-making tasks.
  - Diverse environments and scenarios
  - Comprehensive algorithms support
  - Extensibility, Scalability, Reliability
- **ScalingInter-RL**: a staged training approach that progressively scales agent–environment interactions

Paper

https://agentgym-rl.github.io/

https://github.com/WooooDyy/AgentGym-RL

https://huggingface.co/datasets/AgentGym/AgentGym-RL-Data-ID

Email: zhxi22@m.fudan.edu.cn