Skill Set Optimization: Reinforcing Language Model Behavior via Transferable Skills

UCI AI2

Kolby Nottingham* ◇ Bodhi Majumder ◇ Bhavana Dalvi Mishra
Sameer Singh ◇ Peter Clark ◇ Roy Fox

Motivation

Without costly training, LLM actors have no method to continually adapt and learn
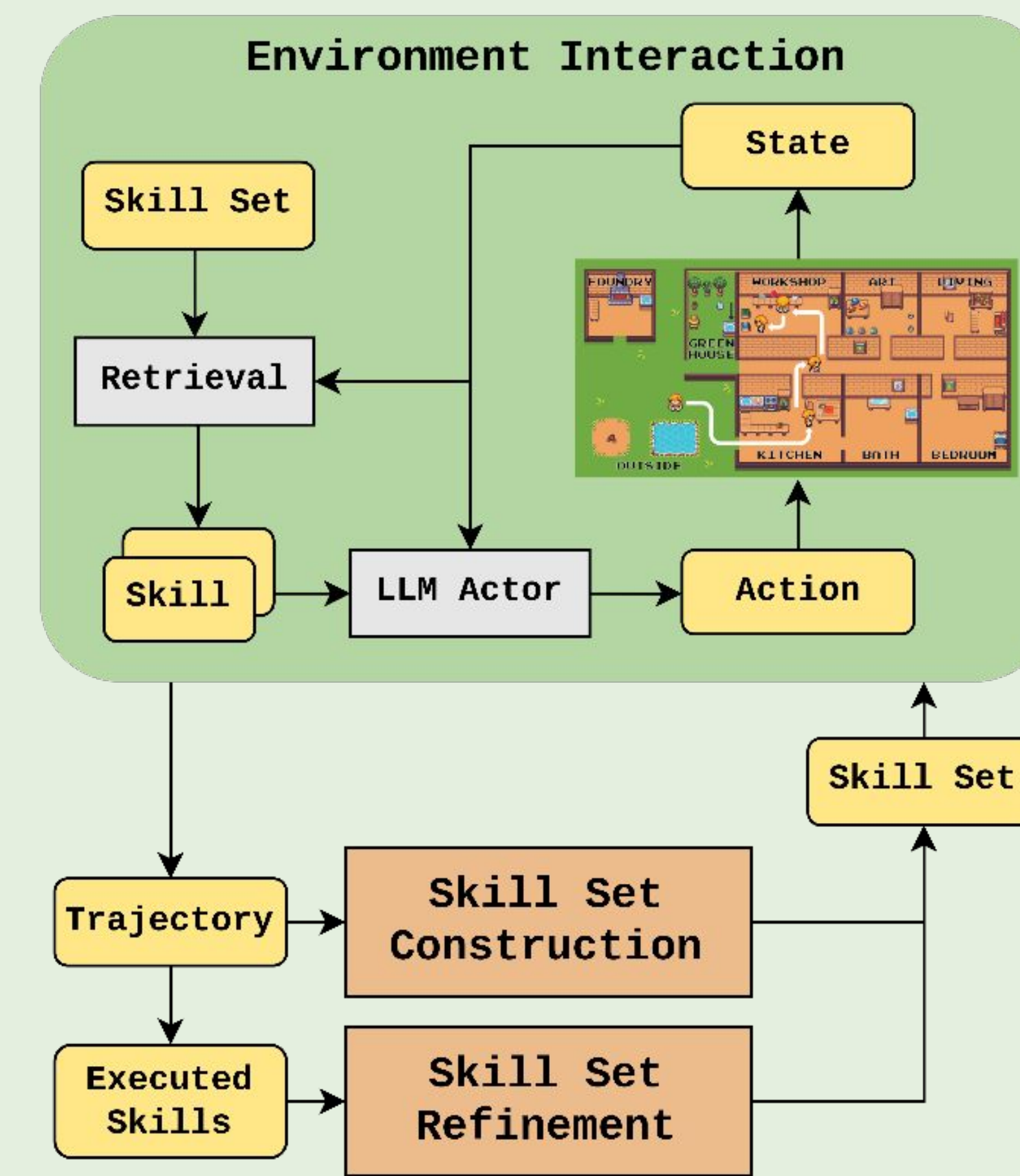
```
Task: Measure the melting temp of chocolate.
State: You see chocolate and a stove.
LLM Actor    move chocolate to stove
State: You see a stove with chocolate.
LLM Actor    activate stove
State: You see a stove with melted chocolate.
LLM Actor    use thermometer on chocolate
State: The thermometer reads 40 C.
```

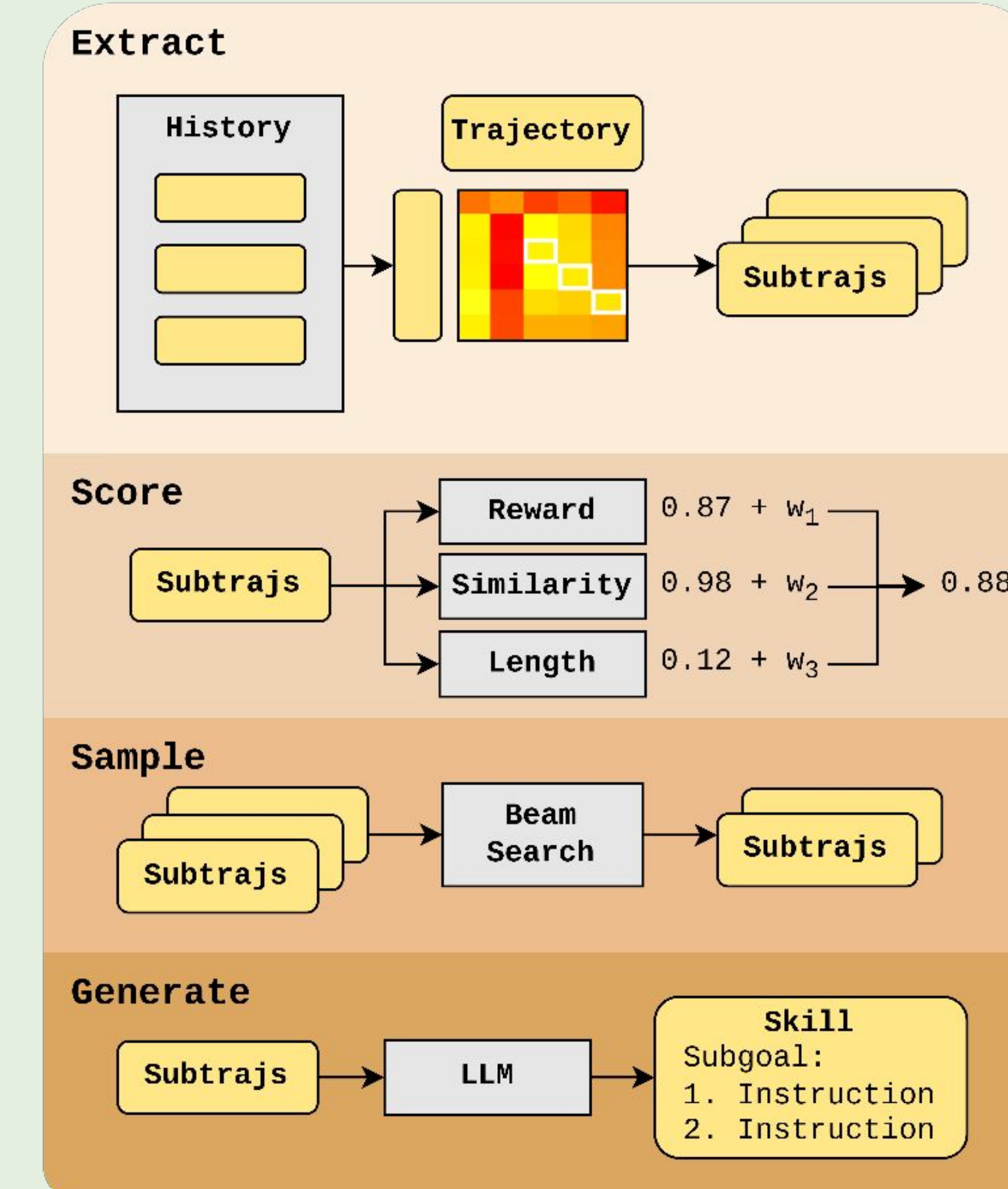SSO summarizes successful experience into retrievable abstract skills

SSO

The unaltered context alone would fail to generalize knowledge to a new task

```
Task: Measure the melting temp of water.
State: You see a stove and a freezer with ice.
```

```
Melt substance:
1. Move substance to stove
2. Activate stove
```

Adding learned skills in-context enables policy improvement

LLM Actor — use thermometer ✗
LLM Actor — move ice to stove ✓

Method

Environment Interaction

Skill Set → Retrieval → Skill → LLM Actor → Action → State

Trajectory → Executed Skills → Skill Set Construction → Skill Set Refinement → Skill Set

Skill Set Optimization occurs after each episode in two steps:

Extract

History → Trajectory → Subtrajs

Score

Subtrajs → Reward $0.87 + w_1$ / Similarity $0.98 + w_2$ / Length $0.12 + w_3$ → 0.88

Sample

Subtrajs → Beam Search → Subtrajs

Generate

Subtrajs → LLM → Skill
Subgoal:
1. Instruction
2. Instruction

1. **Construction** of *subgoals* and *instructions*

Executed Skill → Discounted Reward / Observed Value → (+) → Observed Value → Filter ($\leq \epsilon$ / $> \epsilon$) → Skill
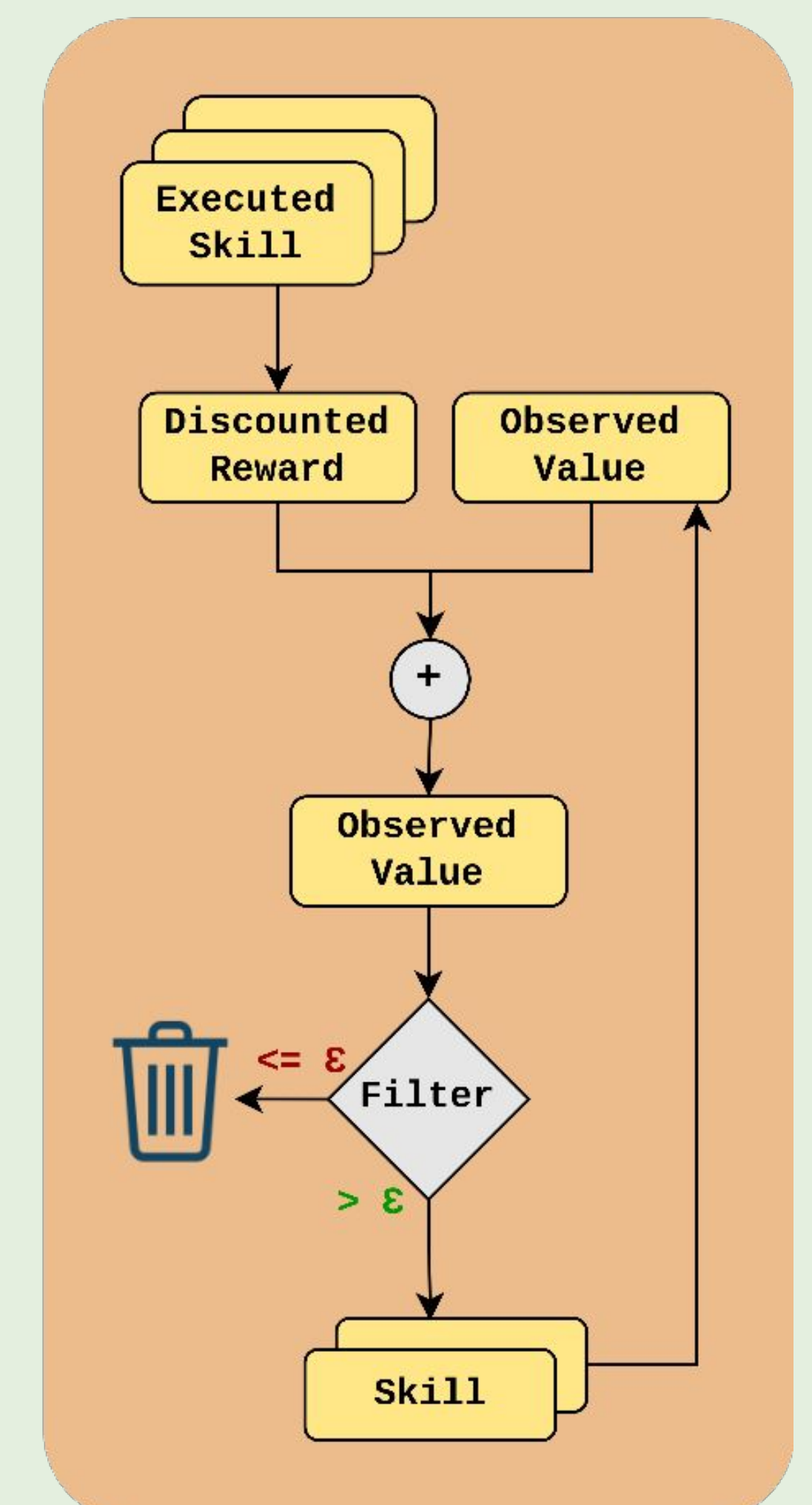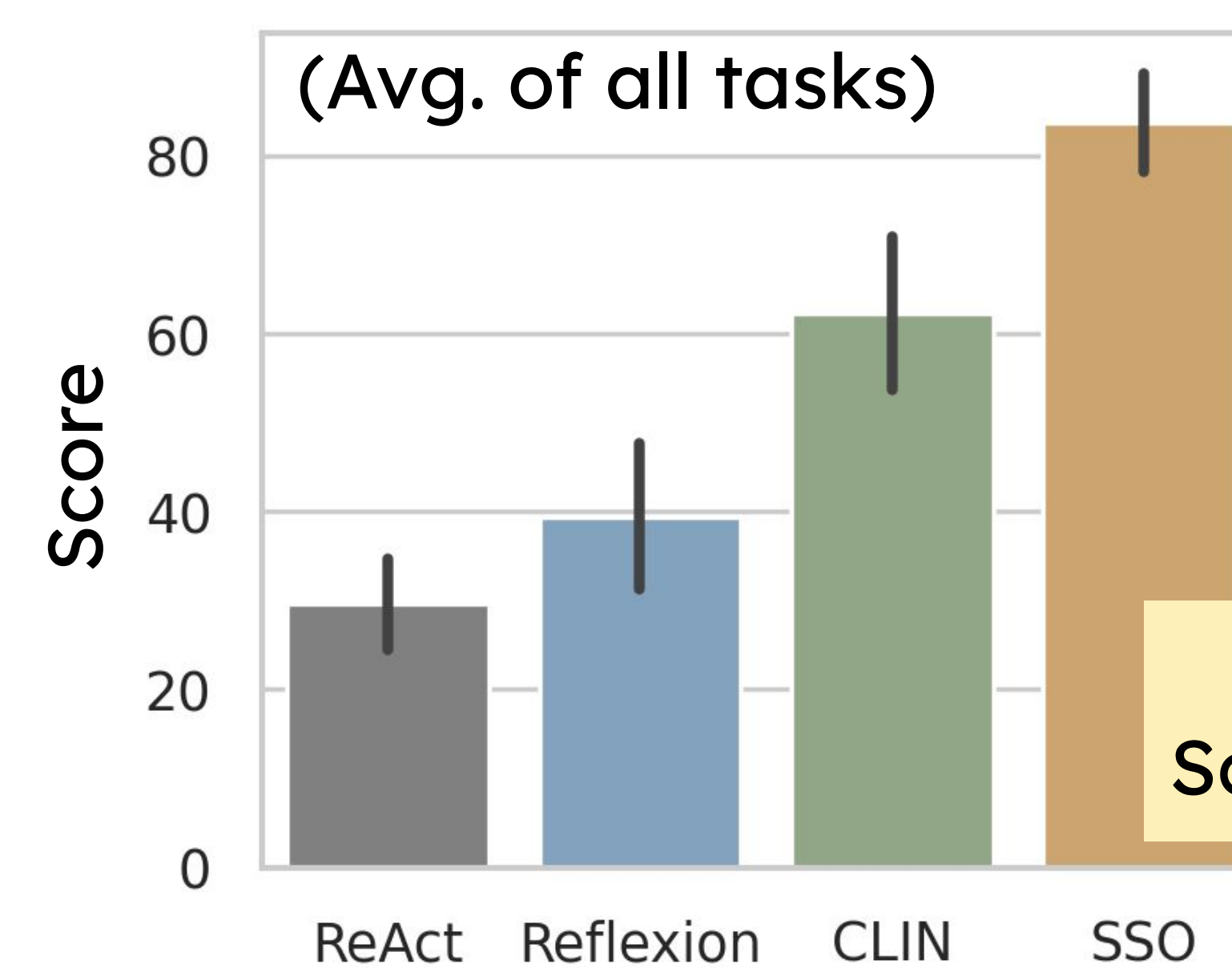
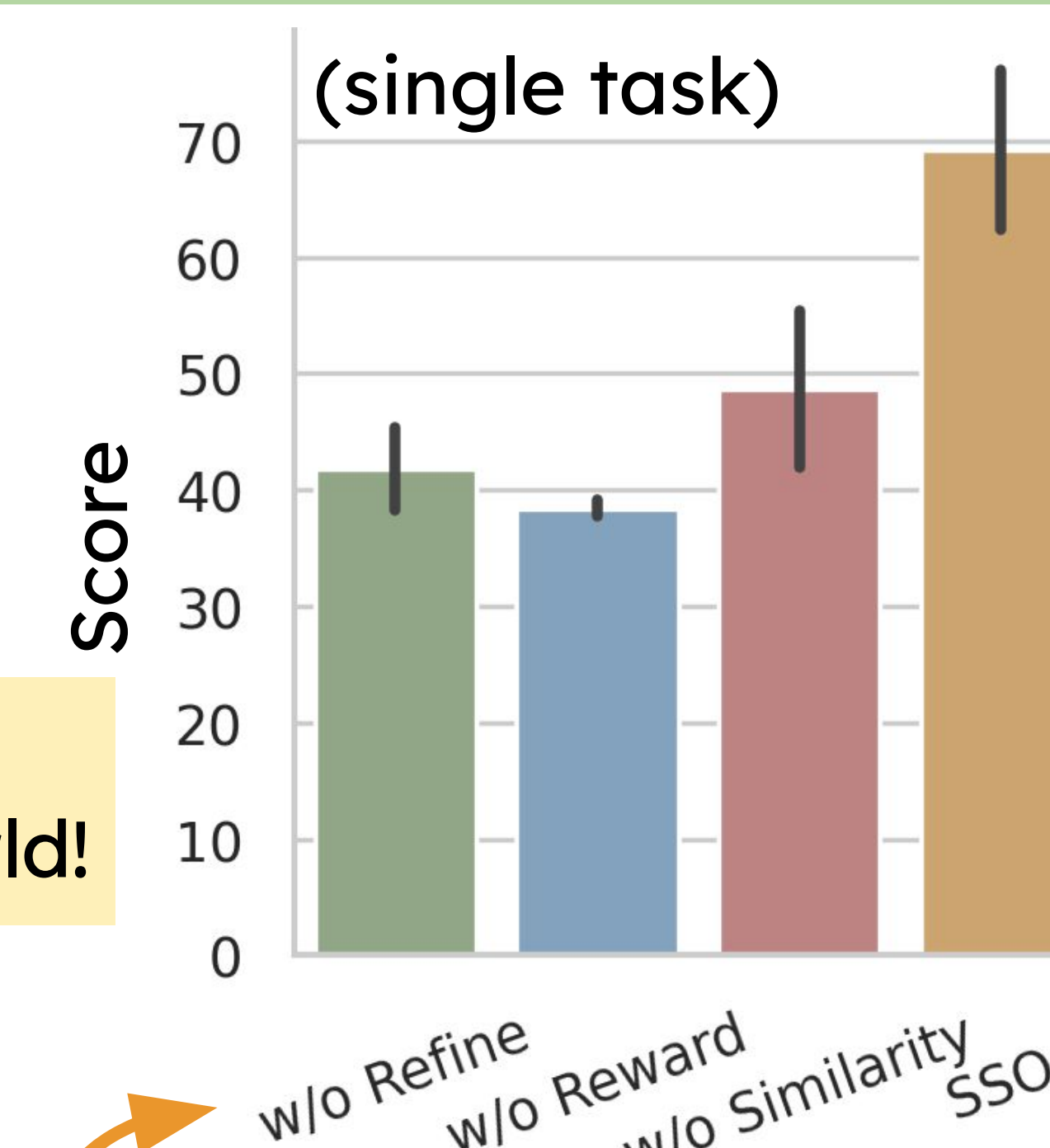2. **Refinement** by pruning poor performing skills

Results

ScienceWorld

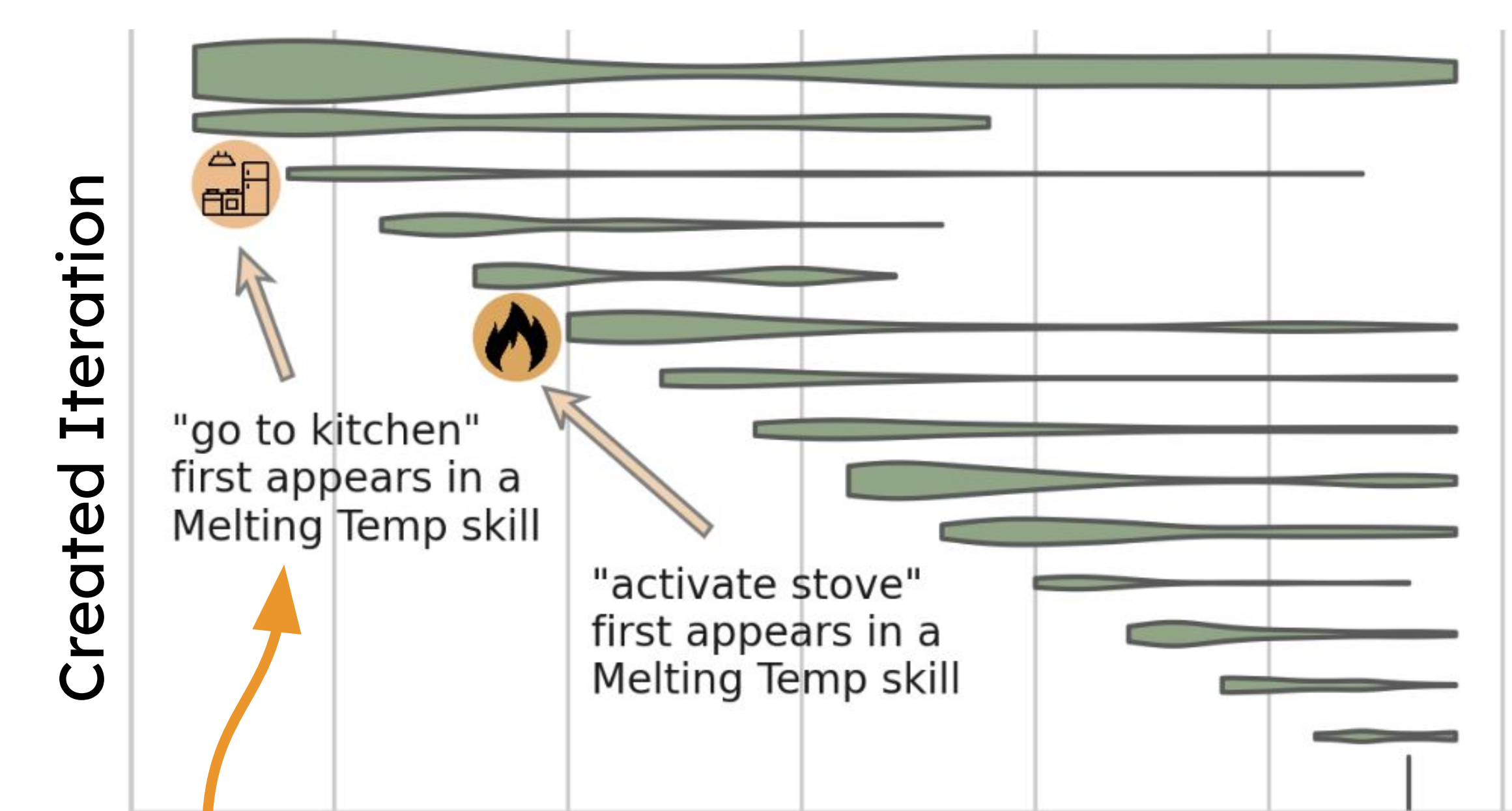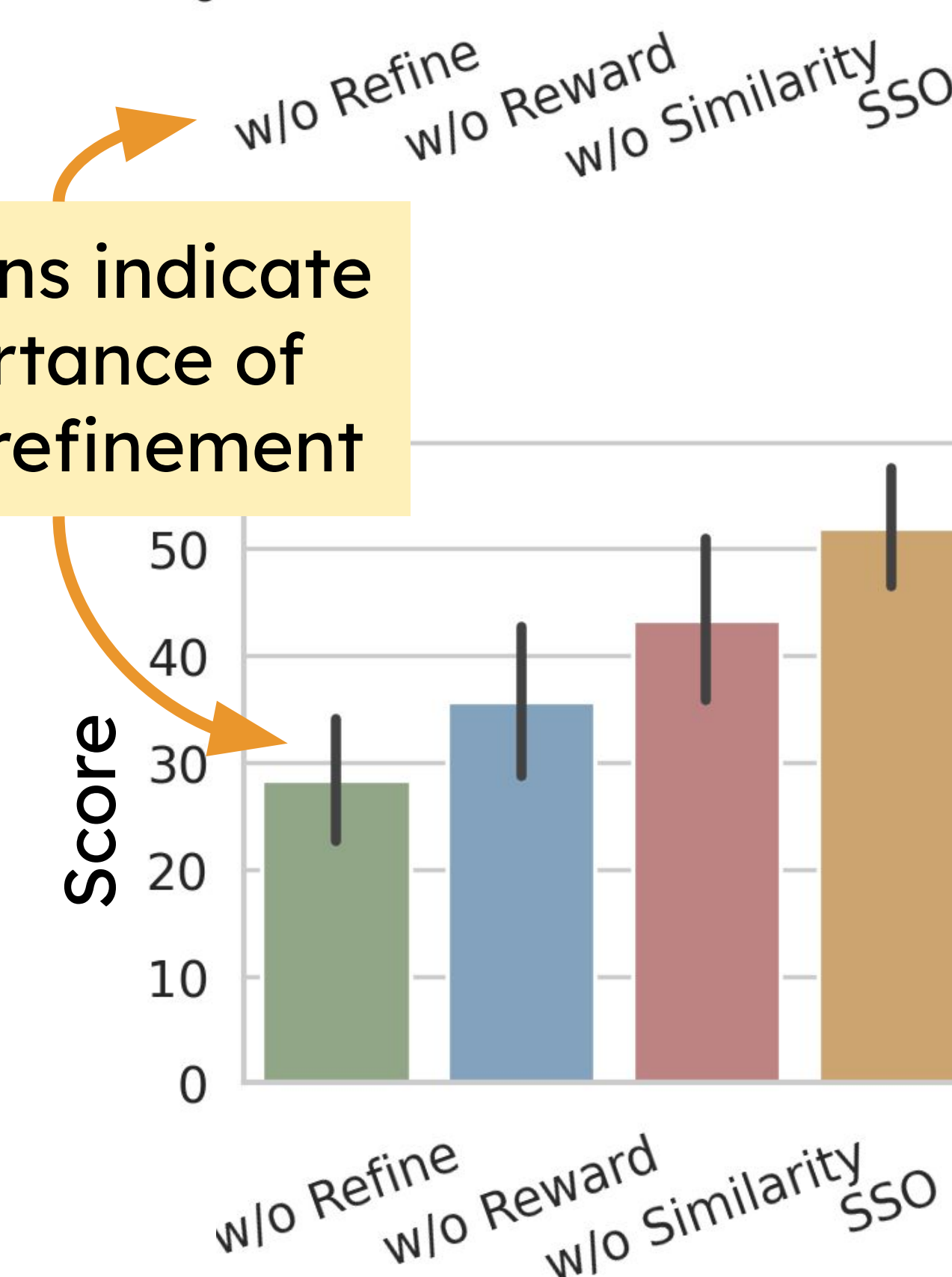Text-based game for simple science experiments and questions. Tests 18 different tasks.

(Avg. of all tasks)

Score: ReAct, Reflexion, CLIN, SSO

SOTA for ScienceWorld!

SSO outperforms in-context baselines

(single task)

Score: w/o Refine, w/o Reward, w/o Similarity, SSO
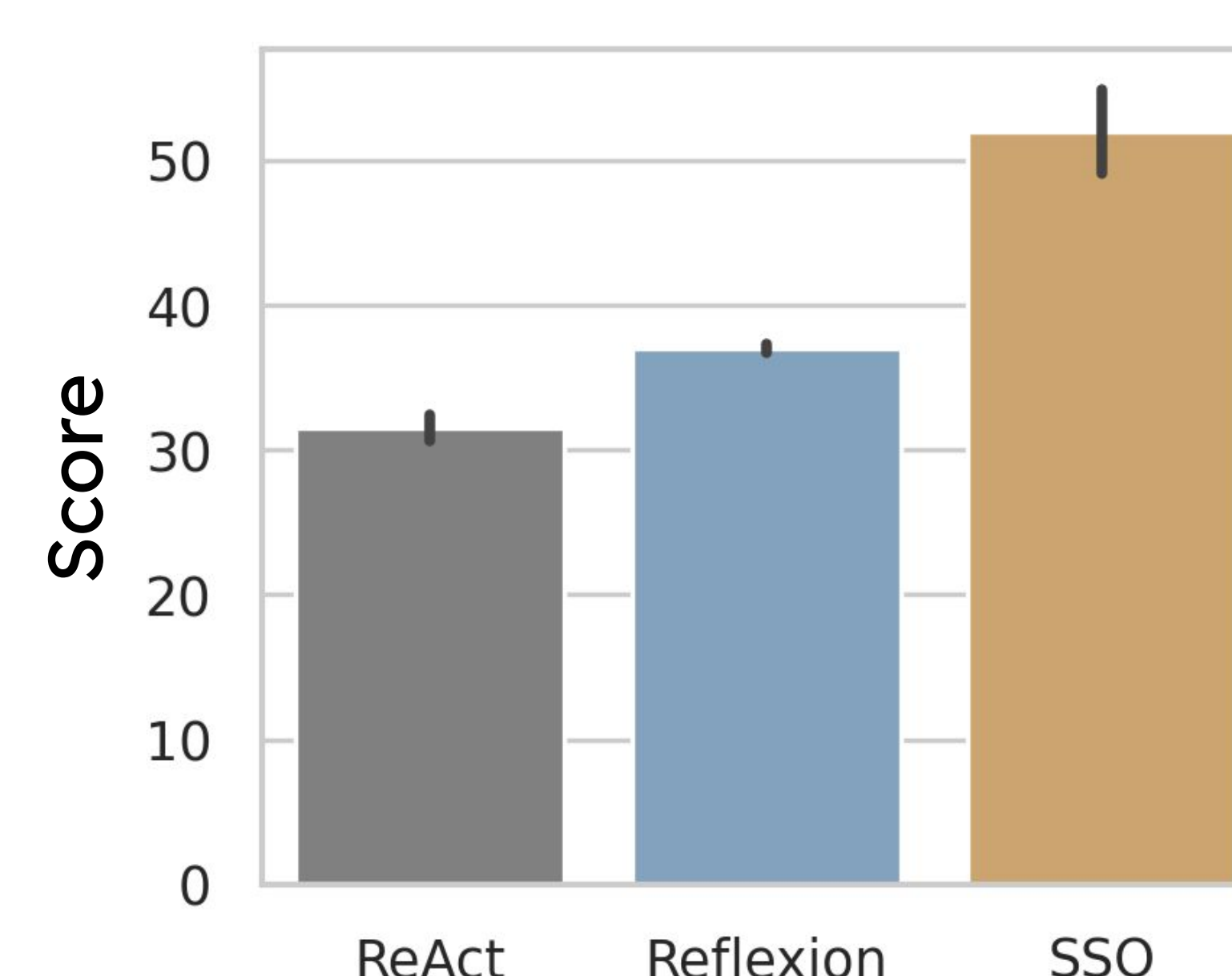
Ablations indicate importance of online refinement

Created Iteration vs Executed Iteration

"go to kitchen" first appears in a Melting Temp skill
"activate stove" first appears in a Melting Temp skill

Skills are learned in order

NetHack

Grid-based game that tests spatial reasoning and object interaction. Agent uses the potion to cross the lava to the stairs.

Score: ReAct, Reflexion, SSO

Score: w/o Refine, w/o Reward, w/o Similarity, SSO

Created Iteration vs Executed Iteration

"pick up key" first appears in a skill
"open door" first appears in a skill

Old skills are forgotten in favor of newer