Experiement Report - Ragen

Prepared by: SIJUN HUA

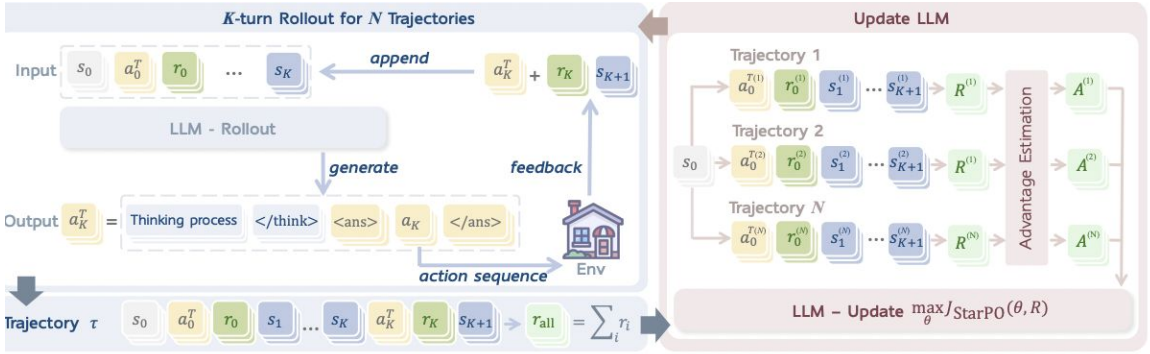
Date: 2025.10.29

System WorkFlow

Explain how your system works overall and integrates A*PO with RAGEN.

- Multi-turn WebShop environment: Search → Inspect → Buy.
- RAGEN framework: Retrieve product information across turns.
- Behavior Cloning warm-start for basic “search → buy” workflow.
- A*PO integrated as the RL optimization core:
  - Uses advantage-weighted policy update
  - KL regularization to stay close to reference model
  - Works well under sparse binary rewards (match / no match)
- Additional reward: Reasoning-Action Consistency bonus (0.0–0.2).

Diagram



User target

Webshop Env

Policy Model

search: green shoe or buy: 42

Trajectory Buffer（stores multiturn traces)

reward compute

A*PO

Basic Settings

Setting	Value
Total Products	**96** (8 colors × 4 sizes × 3 categories)
Colors	black, white, red, blue, green, gray, yellow, brown
Sizes	small, medium, large, xlarge
Categories	shoe, clothing, accessory
Price Range	$20 – $150 (assigned randomly per product)
Available Actions	`search:<query>`, `click:<id>`, `buy:<id>`
Maximum Turns per Episode	5
Search Output	Top-5 matching products returned

Model = qwen2.5-3b
Training Steps = 300
Batch Size = 2
Samples per prompt = 4
learning rate = 3e-5
kl penalty = 0.05
advantage temp = 6.0
top-k = 5

Setting	Value
Fixed Eval Prompts	256 (paper setting)
Evaluation Seed	42 (reproducible)
Max Turns per Episode	5 (truncate interaction)
Success Metric	success = (env_reward == 1.0)
Training-Time Quick Eval	10 cases every 10 steps

Performance

Step	Loss	Avg Reward	Success@10	Notes
10	0.0234	0.05	0.00	Random policy, no task grounding
50	-5.8963	0.18	0.00	Model begins to learn search strategies
100	-17.682	0.34	0.10	Model improves buy decision timing
200	-137.67	0.32	0.10	Policy stabilizing; reasoning aligns with action
300	-420.89	0.40	0.10	Final model converged

Total Advantage = A_env + λ * A_reasoning

where:
- A_env = z_normalize(env_reward) # 0 or 1
- A_reasoning = z_normalize(reasoning) # 0 to 0.2
- λ = 0.4 (reasoning weight)

Final Advantage used in A*PO loss

success = (env_reward == 1.0)

Reasoning score NOT included in success metric

Failure Analysis

Observation:
• The episode reaches the 5-turn limit without executing a “buy” action.

Why This Happens:
• Sparse reward provides no signal until the final step → insufficient exploration.
• Model enters a “search loop,” repeatedly retrieving similar product lists.
• A*PO (α = 6.0) amplifies advantages but still needs initial positive buy examples.

Common Error Patterns:
1. Attribute Prioritization:
   The model prioritizes matching color > category > size > price, which delays commitment.

2. Superficial Reasoning:
   Chain-of-thought often restates the target instead of performing attribute comparison.

3. Late-Stage Skill Dependency:
   Correct buy timing only emerges after Behavior Cloning teaches stable search structure.

THANKS