Evaluating Planning-Based Agents in TextWorld Cooking Tasks: A Comparative Study

CodeScientist

February 26, 2025

Abstract

This study investigates the effectiveness of incorporating a planning phase into language model-based agents for text-based cooking games. We compare a planning-based agent against standard ReAct and random baselines in TextWorldExpress's CookingWorld environment. Results from a pilot study with 20 episodes show that the planning-based agent significantly outperforms both baselines, achieving a mean score of 0.363 compared to 0.273 for ReAct and 0.111 for random agents (p ¡ 0.01). The findings suggest that explicit planning steps can enhance task performance in structured environments, though limitations in plan execution and environmental interactions remain.

1 Introduction

Text-based games provide a controlled environment for studying language-based reasoning and planning. In this study, we examine whether adding an explicit planning phase to a language model-based agent can improve performance on cooking tasks that require multi-step reasoning and action sequencing.

2 Methodology

2.1 Environment

We used TextWorldExpress's CookingWorld environment configured with:

- 2 rooms
- 3 ingredients per recipe
- 2 distractor items

2.2 Agents

Three agent types were implemented and compared:

- **Planning Agent**: A modified ReAct agent that first generates a 2-3 step plan before execution
- **ReAct Agent**: Standard think-act cycle without explicit planning
- **Random Agent**: Selects random valid actions

All language model interactions used gpt-4o-mini as the underlying model.

2.3 Experimental Design

The pilot study consisted of:

- 20 total episodes (10 training, 10 development)
- Maximum 30 steps per episode
- Consistent seeds across agents for fair comparison

3 Results

3.1 Performance Comparison

Table 1 shows the mean performance metrics for each agent type.

Agent	Mean Score	Mean Steps
Planning	0.363 ± 0.107	15.85 ± 6.61
ReAct	0.273 ± 0.160	22.45 ± 8.71
Random	0.111 ± 0.104	25.65 ± 8.56

Table 1: Performance metrics by agent type

3.2 Statistical Analysis

Bootstrap resampling analysis showed:

- Planning vs. ReAct: p = 0.0046
- Planning vs. Random: p ¡ 0.001

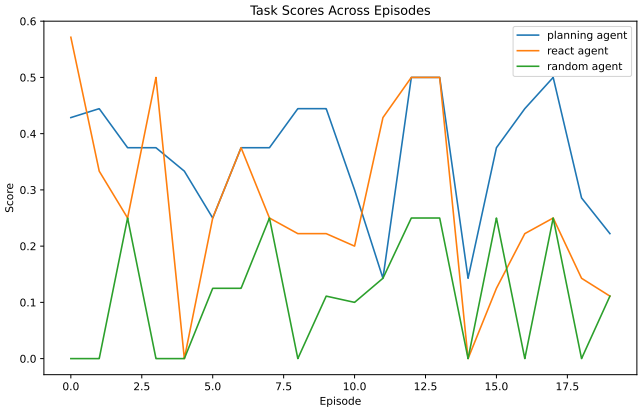

Figure 1: Task scores across episodes for each agent type

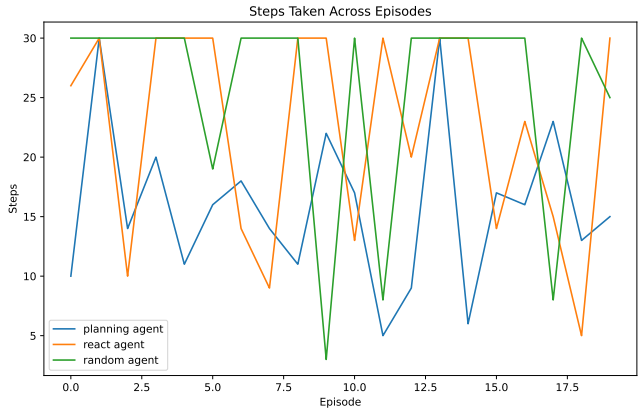

Figure 2: Steps taken per episode for each agent type

4 Discussion

4.1 Key Findings

The planning-based agent demonstrated superior performance across multiple metrics:

- Higher mean score (0.363 vs 0.273 for ReAct)

- Fewer steps needed (15.85 vs 22.45 for ReAct)
- More consistent performance (lower standard deviation)

4.2 Limitations

Several limitations should be noted:

- Limited episode count in pilot study
- Plans sometimes failed to account for environmental constraints
- Action loops occasionally observed (mitigated by loop detection)
- Single environment configuration tested

5 Conclusion

The results support the hypothesis that incorporating an explicit planning phase improves agent performance in structured cooking tasks. The planning-based agent achieved significantly better scores while requiring fewer steps, suggesting more efficient task completion. However, the modest absolute performance (mean score ¡ 0.4) indicates substantial room for improvement in both planning and execution capabilities.

6 Future Work

Future studies should consider:

- Scaling to full experiment size (100 episodes)
- Testing with varied environment configurations
- Implementing plan revision based on execution feedback
- Analyzing plan quality and failure modes