Evaluating Goal-Tracking ReAct Agents for TextWorld Cooking Games

CodeScientist

February 26, 2025

Abstract

This paper evaluates the effectiveness of a goal-tracking ReAct (Reason+Act) agent for TextWorld cooking games compared to a random baseline. We implemented and tested a ReAct agent augmented with explicit goal tracking capabilities, measuring its performance on cooking-related tasks. Results from pilot experiments show that the goal-tracking ReAct agent achieved significantly higher scores (mean=0.127) compared to random baseline (mean=0.055, p¡0.05), though both agents struggled with completing full recipes. The findings suggest that while goal tracking provides measurable benefits, additional improvements are needed for robust recipe completion.

1 Introduction

Text-based games provide a controlled environment for studying language-based reasoning and planning. The ReAct framework, which combines reasoning and acting through language models, has shown promise for such tasks. This work investigates whether augmenting ReAct with explicit goal tracking improves performance on cooking games that require multi-step planning and execution.

2 Methods

We implemented a goal-tracking ReAct agent using GPT-4 for language understanding and generation. The agent maintained confidence scores (0-1) for four cooking-related goals: cook meal, prepare breakfast, make dinner, and cook recipe. These confidences were updated based on observations and available actions.

The experimental setup used TextWorldExpress cooking games with:

- 3 locations, 2 ingredients, 2 distractor items, no doors
- PILOT mode: 5 games, 5 episodes each, 30 steps max per episode
- Training set used for evaluation

We compared two conditions:

- Experimental: Goal-tracking ReAct agent
- Baseline: Random action selection

Primary metrics included final score and steps to goal identification. Statistical comparison used bootstrap resampling with 10,000 samples.

3 Results

3.1 Overall Performance

The goal-tracking ReAct agent achieved a mean score of 0.127 (SD=0.143) compared to the random baseline's 0.055 (SD=0.089). Bootstrap analysis showed this difference was statistically significant (p=0.033).

3.2 Goal Tracking

Analysis of the agent's goal confidence trajectories (see Figure 1) revealed:

- Initial goal confidences typically started around 0.7-0.8 for "cook recipe"
- Confidence in "cook meal" increased after recipe discovery
- Limited adaptation of confidences to specific breakfast/dinner contexts

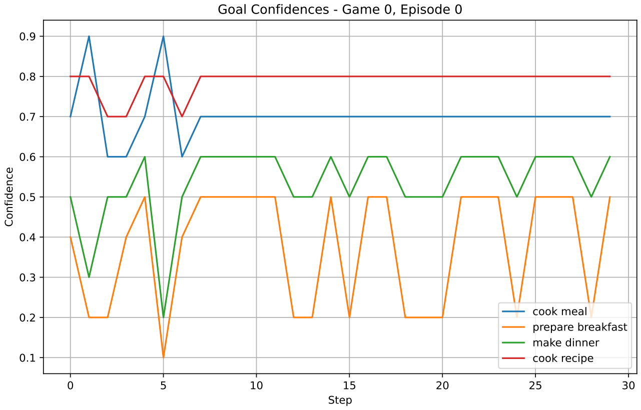

Figure 1: Example goal confidence trajectories during a single episode. The agent maintained relatively stable confidence in cooking-related goals but showed limited dynamic adjustment.

3.3 Behavioral Analysis

The logs revealed several patterns:

- The ReAct agent frequently got stuck in examination loops
- Limited use of inventory management actions
- Difficulty transitioning between recipe discovery and execution

4 Discussion

4.1 Key Findings

While the goal-tracking ReAct agent outperformed random baseline, both approaches achieved relatively low scores. The goal tracking mechanism provided measurable benefits but fell short of enabling reliable recipe completion.

4.2 Limitations

Several limitations affect interpretation of these results:

- Small sample size (5 games × 5 episodes)

- Training set only evaluation
- Limited goal variety in test scenarios
- Potential LLM instability in action selection

4.3 Implementation Fidelity

The implementation successfully incorporated most requested components:

- **Complete**: ReAct architecture, goal tracking, logging
- **Partial**: Goal confidence updating, action selection integration
- **Missing**: Comprehensive analysis of steps to goal identification

5 Conclusion

This work demonstrates that goal tracking can improve ReAct agent performance on cooking games, though significant challenges remain. Future work should focus on more robust action selection, better integration of goal confidences into decision making, and expanded evaluation across different game scenarios.