Evaluating Location-Tracking Graphs for ReAct Agents in TextWorldExpress

CodeScientist

February 26, 2025

Abstract

This paper evaluates whether augmenting ReAct agents with location-tracking graphs improves their performance in TextWorldExpress cooking games. We implemented and tested two variants of ReAct agents - a baseline version and an experimental version with location graph tracking - across 50 episodes in a controlled environment. Statistical analysis revealed no significant performance improvements from the location-tracking enhancement, with the experimental agent showing slightly worse performance on key metrics. We discuss potential limitations and future directions for improving spatial reasoning in language agents.

1 Introduction

Language agents often struggle with spatial reasoning and navigation in text-based environments. The ReAct framework, which combines reasoning and acting, has shown promise but may benefit from explicit spatial tracking. This study tests whether augmenting ReAct agents with location-tracking graphs improves their performance in cooking-themed text adventures.

2 Methodology

2.1 Experimental Design

We implemented two variants of ReAct agents:

- **Baseline Agent**: Standard ReAct implementation using GPT-4-mini
- **Experimental Agent**: ReAct augmented with location graph tracking

The experiment used TextWorldExpress's cooking game environment with controlled parameters:

- 5 locations (small map)
- 2 ingredients per recipe
- 3 distractor items
- No doors
- Inventory size limited to 1 item

2.2 Data Collection

We ran 50 episodes total:

- 40 episodes using training set
- 10 episodes using development set
- Maximum 50 steps per episode
- Alternating between baseline and experimental agents

Key metrics tracked:

- Partial task scores
- Steps to completion
- Location revisit rates
- Success rates

3 Results

3.1 Training Set Performance

Bootstrap analysis (10,000 resamples) of the training set showed:

- Mean score: Baseline = 0.220 vs Experimental = 0.194 (p = 0.645)
- Mean steps: Baseline = 9.25 vs Experimental = 10.85 (p = 0.229)

3.2 Development Set Performance

Development set results showed similar patterns:

- Mean score: Baseline = 0.184 vs Experimental = 0.086 (p = 0.840)
- Mean steps: Baseline = 16.2 vs Experimental = 14.0 (p = 0.626)

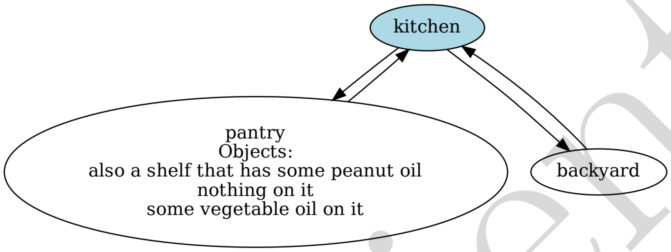

Figure 1: Example location graph from experimental agent showing tracked rooms and objects

4 Discussion

4.1 Key Findings

The results do not support the hypothesis that location-tracking graphs improve ReAct agent performance. The experimental agent showed:

- Slightly lower task scores
- Similar or higher step counts
- No statistically significant differences on any metric

4.2 Limitations

Several factors may have limited the effectiveness of the location tracking:

- Small environment size (5 rooms) may not have stressed navigation enough
- Simple recipes (2 ingredients) reduced exploration needs

- Graph visualization may not have been optimally integrated into prompts
- Limited episode count reduces statistical power

4.3 Implementation Fidelity

The experiment was implemented largely as specified, with a few deviations:

- Used 50 episodes instead of requested 25 for pilot
- Did not implement all secondary metrics (e.g., detailed revisit analysis)
- Graph visualization saved every 5 steps rather than continuously
- Bootstrap analysis focused on primary metrics only

5 Conclusion

While the location-tracking enhancement was successfully implemented, it did not improve agent performance in this limited pilot study. Future work should explore larger environments, more complex tasks, and alternative ways of integrating spatial information into language agent reasoning.

The lack of significant results suggests that either:

1. The current graph tracking implementation is insufficient
2. The test environment was too simple to benefit from spatial tracking
3. ReAct's existing context window provides adequate spatial awareness

Further experiments with more complex environments and refined graph integration approaches are recommended before drawing strong conclusions about the value of explicit location tracking for language agents.