Evaluating LLM-based Verification for Knowledge Graph Accuracy in Text-Based Games

CodeScientist

February 26, 2025

Abstract

This paper evaluates whether Large Language Model (LLM) verification can improve the accuracy of automatically extracted knowledge graphs in text-based game environments. Using a controlled experiment in the CookingWorld domain, we compare knowledge graphs generated with and without LLM-based triple verification. Results from a pilot study (n=10 episodes) show that LLM verification significantly improves graph quality, with an 81.9% verification rate and higher game performance scores (p ¡ 0.001). However, limitations in sample size and game complexity suggest the need for larger-scale validation.

1 Introduction

Knowledge graphs are crucial for representing game state in text-based environments, but automatic extraction from natural language descriptions is error-prone. This study tests whether LLM-based verification can improve knowledge graph accuracy by validating extracted triples.

2 Methods

2.1 Experimental Design

We implemented a controlled experiment comparing two conditions:

- Baseline: Direct knowledge graph generation without verification
- Experimental: Knowledge graph generation with LLM-based triple verification

The experiment used CookingWorld from TextWorldExpress with parameters: numLocations=3, numIngredients=2, numDistractorItems=2, includeDoors=0. We ran 10 episodes of 30 steps each (PILOT mode).

2.2 LLM Verification Process

For the experimental condition, each extracted triple was verified using gpt-4o-mini with the prompt:

```
Given the game state description: '{state_desc}'
Is the following knowledge graph triple correct? '{triple}'
Respond in JSON format with keys 'is_correct' (boolean)
and 'correction' (string, only if is_correct is false).
```

3 Results

3.1 Verification Performance

The LLM verification system demonstrated strong performance:

- Total triples verified: 1,108
- Correctly verified: 875 (79.0%)
- Incorrect/corrected: 233 (21.0%)
- Average verification rate: 79.5%

3.2 Graph Metrics

Table 1 shows the average graph metrics across conditions.

Table 1: Average Graph Metrics by Condition

Metric	Baseline	With Verification
Nodes	40.2	32.5
Edges	54.5	44.3
Density	0.034	0.042

3.3 Game Performance

Bootstrap analysis (10,000 resamples) showed significantly better game performance with LLM verification:

- Baseline mean score: 0.067
- Experimental mean score: 0.152
- p-value: 0.001

4 Discussion

4.1 Key Findings

The results support the hypothesis that LLM-based verification improves knowledge graph quality. Key evidence includes:

1. High verification rate (79.5%) indicating effective error detection
2. More concise graphs (fewer nodes/edges) suggesting noise reduction
3. Significantly better game performance with verification

4.2 Limitations

Several limitations should be noted:

1. Small sample size (10 episodes) limits statistical power
2. Simple game environment may not generalize to more complex scenarios
3. No ground truth validation of graph accuracy
4. Potential LLM hallucination effects not controlled for

5 Conclusion

This pilot study provides preliminary evidence that LLM-based verification can improve knowledge graph accuracy in text-based games. The significant improvement in game performance suggests that verified graphs better capture relevant game state information. However, larger-scale studies with more complex environments are needed to validate these findings.

6 Future Work

Future research directions include:

- Scaling to full experiment size (100 episodes)
- Testing in more complex game environments
- Comparing different LLM models for verification
- Adding ground truth validation of graph accuracy