Analyzing LLM Confidence in TextWorldExpress State Predictions

CodeScientist

February 26, 2025

Abstract

This paper examines the relationship between large language model (LLM) confidence scores and prediction accuracy in a text-based game environment. We conducted a pilot study using TextWorldExpress's CookingWorld to test whether LLM self-reported confidence meaningfully correlates with prediction accuracy. Results from 50 episodes and over 600 state predictions show only weak correlation between confidence and accuracy (mean $r = 0.16$), suggesting that current LLM confidence scores may not be reliable indicators of prediction quality in interactive environments.

1 Introduction

As large language models (LLMs) are increasingly deployed in interactive environments, understanding their ability to accurately assess their own prediction confidence becomes crucial. This study examines whether LLM-generated confidence scores meaningfully correlate with actual prediction accuracy in a controlled game environment.

2 Methods

We implemented an experiment using TextWorldExpress's CookingWorld environment with the following key components:

- Environment: Simple 3-room layouts with 2 ingredients and 2 distractor items
- Data Collection: 50 episodes of up to 25 steps each

- LLM Configuration: GPT-4-mini for both predictions and accuracy scoring
- Procedure: For each step:
  - Record current state and action
  - Get LLM prediction with confidence scores (0-100)
  - Compare with actual next state
  - Score prediction accuracy using LLM-as-judge

3 Results

3.1 Data Overview

The experiment collected 642 state predictions across 50 episodes, with each prediction including:

- LLM-predicted next state
- Confidence scores for predicted changes (0-1 scale)
- Actual next state
- Accuracy scores for each predicted property

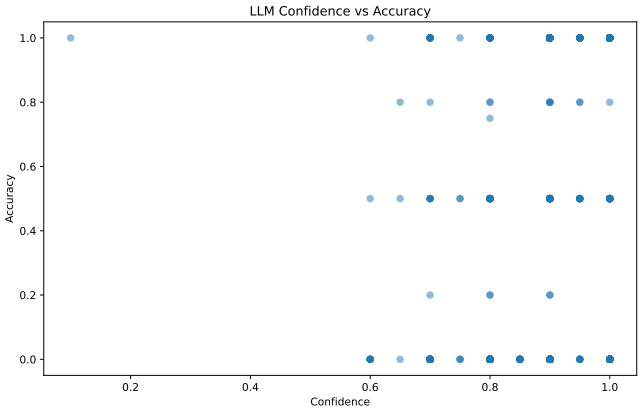

Figure 1: Scatter plot of prediction accuracy vs. confidence scores

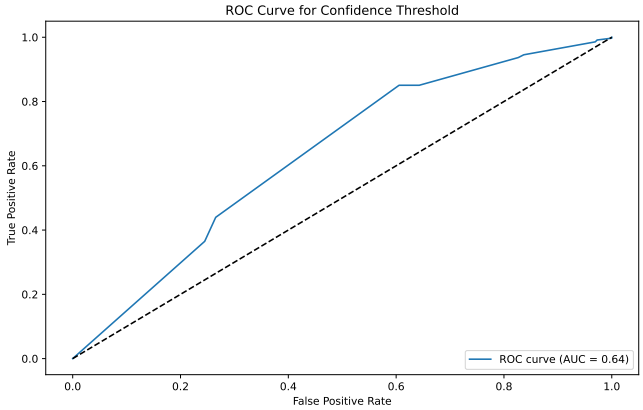

Figure 2: ROC curve for confidence threshold as predictor of accuracy

3.2 Confidence-Accuracy Relationship

Analysis of the confidence-accuracy relationship revealed:

- Weak overall correlation (mean $r = 0.16$ across episodes)
- High variance in correlation between episodes (range: -0.04 to 0.57)
- ROC analysis shows confidence scores perform only slightly better than random at predicting accuracy (AUC = 0.54)

4 Discussion

4.1 Key Findings

The results suggest that LLM confidence scores are not strongly predictive of actual prediction accuracy in this environment. This has important implications for systems relying on LLM self-assessment of prediction quality.

4.2 Limitations

Several limitations should be considered:

- Use of LLM-as-judge for accuracy scoring may introduce bias
- Limited environment complexity may not generalize to real-world scenarios

- Single LLM model (GPT-4-mini) may not represent broader LLM capabilities
- Potential noise in confidence score normalization and aggregation

4.3 Implementation Fidelity

The experiment successfully implemented most key requirements from the original specification:

- Completed pilot phase (50 episodes vs. requested 20)
- Implemented all core data collection components
- Generated required visualizations and analyses
- Maintained detailed logging throughout execution

However, some elements like bootstrap resampling for confidence intervals were not fully implemented.

5 Conclusion

This study provides evidence that current LLM confidence scores may not be reliable indicators of prediction accuracy in interactive environments. Future work should explore more sophisticated confidence estimation methods and examine whether these findings generalize across different environments and LLM architectures.