TOWARD TEMPLATE-FREE EXPLAINABILITY FOR MONTE CARLO TREE SEARCH

Siqi Lu[1], Mirsaleh Bahavarnia[2], Hiba Baroud[2], Yixuan Zhang[1], Hemant Purohit[3], Ayan Mukhopadhyay[1*]

[1] William & Mary · [2] Vanderbilt University · [3] George Mason University

*Corresponding Author: amukhopadhyay@wm.edu

WILLIAM & MARY

IJCAI-ECAI 2026 XAI Workshop

The Challenge

Monte Carlo Tree Search (MCTS) have proven very effective in solving sequential decision-making tasks under uncertainty.

However, its decisions are usually difficult for end users to interpret because they are derived from large, asymmetric search trees, stochastic simulations, and low-level statistics.

Existing explanation methods can produce grounded explanations, but they rely on hand-crafted logic templates and domain-specific mapping that must be redesigned whenever the task or environment changes.

→ How can we explain MCTS decisions accurately and flexibly--without hand-crafted, domain-specific templates?

Contributions

01 – USER-CENTERED EXPLAINABILITY

Our method enables users to ask natural-language questions about the reasoning behind MCTS decisions—not just inspect final actions or low-level statistics.

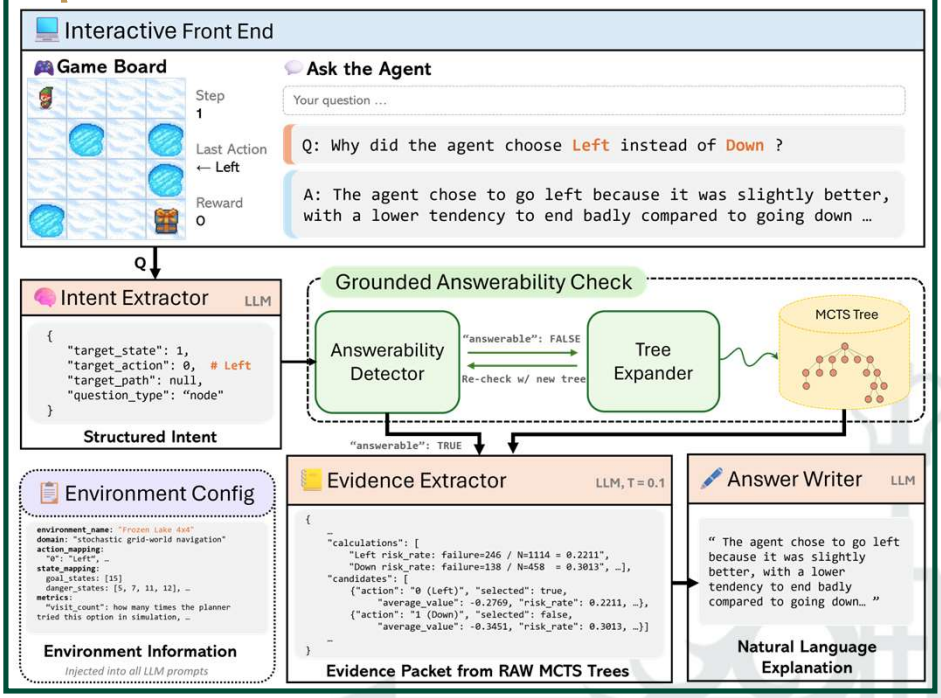
02 – END-TO-END & EVIDENCE-GROUNDED FRAMEWORK

We propose an interactive LLM-based end-to-end framework that connects user questions with raw MCTS traces and generates explanations grounded in the search process. *To the best of our knowledge, this is the first framework that enables the explainability of probabilistic search trees without any intermediate formal representation.*

03 - SUPPORT FOR DIVERSE USER QUESTIONS

This method supports *why, why-not,* and *what-if* questions across different explanation scopes.

An End-to-End Explanation Framework



Evaluation

We evaluate the framework on 21 annotated natural-language questions in the Frozen Lake environment.

The questions cover:

- Question forms: why, why-not, and what-if
- Explanation scopes: node-, path-, and general-level
- Evidence conditions: answerable and unanswerable

The evaluation results are in table 1, 2, and 3.

Key Findings

- ✓ The intent extractor correctly identify all the targeted structured intents.
- ✓ The answerability detector correctly classified all 21 queries.
- ✓ 99.1% of extracted metric values matched the ground truth recomputed from the MCTS traces.
- ✓ All 150 calculations used the correct metric formulas.
- ✓ Remaining errors primarily involved incomplete evidence coverage, rather than incorrect calculations.

Qualitative Results

Evaluation Environment: Frozen Lake

WHY

Q: Why did the agent choose Left instead of Down?

A: The agent chose Left instead of Down because it tended to result in better outcomes, with about 96% of the attempts ending back at the current state and only about 22% of the attempts ending in a hole, compared to Down which had roughly 27% of attempts ending in a hole.

WHY-NOT

Q: Why did the agent not choose Down?

A: The agent chose Left over Down because it tended to end slightly better, with a lower risk of falling into a hole, about 22% of the time for Left, compared to about 27% for Down.

WHAT-IF

Q: What if the agent chose Down instead?

A: If the agent chose Down instead, it would have succeeded about 1% of the time and failed roughly 27% of the time, making it clearly worse than the chosen action.

ENVIRONMENT-RELATED

Q: What is the slipperiness of this game?

A: The agent has a 10% chance of slipping sideways, with a 90% chance of moving in the intended direction.

Table 1: Intent extraction results on the annotated query set.

Metric	Correct / Total	Accuracy
Question type	21 / 21	100.0%
Target state	21 / 21	100.0%
Target action	21 / 21	100.0%
Target path	21 / 21	100.0%

Table 2: Answerability detection results on the annotated query set.

Metric	Correct / Total	Accuracy
Answerability decision	21 / 21	100.0%

Table 3: Evidence-correctness results on the annotated query set.

Check	Passed / Total	Rate
Coverage (per query)	17 / 21	81.0%
Value fidelity (per query)	19 / 21	90.5%
Metric values (per value)	335 / 338	99.1%
Formula correctness (per calc.)	150 / 150	100.0%
All checks passed (per query)	17 / 21	81.0%

Funding and Acknowledgements