Detailed Report on Agent Testing and Evaluation

Table of Contents

1. Introduction
2. Initial Testing of the Agent
    - Test with 10 Questions
    - Results and Observations
3. Re-testing with Hints and Requesting Reasoning Steps
    - Test with 10 Questions and Hints
    - Requesting Detailed Reasoning Steps
    - Results and Observations
4. Review and Critique of Agent’s Reasoning
    - Agent’s Self-Review of Prior Reasoning
    - Identification and Critique of Errors
5. Additional Tasks Assigned to the Agent
    - Task to Research and Create Strategy Guide
    - Agent’s Performance and Observations
6. Analysis of Agent’s Capabilities
    - Strengths
    - Weaknesses
7. Recommendations and Future Work
8. Conclusion

Introduction

This report documents the comprehensive testing and evaluation of an AI agent’s capabilities in mathematical problem-solving and task execution. The testing process involved a series of stages: 1. Initial testing with 10 mathematical problems to assess baseline performance. 2. Re-testing with hints provided for the same problems, requesting detailed reasoning steps. 3. Review and critique of the agent’s reasoning, identifying errors and areas for improvement. 4. Assignment of additional tasks, including researching and compiling a strategy guide.

The goal was to evaluate the agent’s ability to solve complex problems, incorporate feedback, improve reasoning, and perform tasks requiring research and critical thinking.

Initial Testing of the Agent

Test with 10 Questions

The agent was presented with a set of 10 mathematical problems of varying difficulty levels, primarily sourced from mathematical olympiads and advanced competitions. The problems covered topics such as number theory, geometry, combinatorics, and algebra.

Example Problem: 1. Three airline companies operate flights from Dodola island. Each company has a different schedule of departures. The first company departs every 100 days, the second every 120 days, and the third every 150 days. What is the greatest positive integer for which it is true that there will be consecutive days without a flight from Dodola island, regardless of the departure times of the various airlines?

Results and Observations

- Correct Answers: The agent provided correct answers to some of the problems, demonstrating a basic understanding of mathematical concepts.
- Incorrect Answers: For several problems, the agent either provided incorrect answers or failed to provide a solution.
- Lack of Reasoning: The agent's responses lacked detailed explanations or step-by-step reasoning, making it difficult to assess the thought process behind the answers.
- Areas of Difficulty: The agent struggled particularly with problems requiring advanced problem-solving techniques or deeper mathematical insights.

Re-testing with Hints and Requesting Reasoning Steps

Test with 10 Questions and Hints

To assist the agent, hints were provided for each of the 10 problems. The hints aimed to guide the agent toward the correct solution path without explicitly giving away the answer.

Example Problem with Hint: 1. Hint: Consider calculating the least common multiple (LCM) of the departure intervals to determine the cycle of departures.

Requesting Detailed Reasoning Steps

In addition to providing hints, the agent was explicitly instructed to include detailed reasoning steps in the solutions. The purpose was to encourage the agent to break down the problems and demonstrate the logical processes leading to the answers.

Results and Observations

- Improved Accuracy: With hints, the agent showed improvement in solving the problems, providing more correct answers than in the initial test.
- Inclusion of Reasoning Steps: The agent began to include step-by-step explanations, outlining the methods used to reach the solutions.
- Variable Depth of Reasoning: The depth and clarity of reasoning varied across problems. Some solutions were well-explained, while others remained superficial.
- Reliance on Hints: The agent's improved performance highlighted a dependency on the provided hints, indicating challenges in independent problem-solving.

Review and Critique of Agent's Reasoning

Agent's Self-Review of Prior Reasoning

The agent was tasked with reviewing its previous solutions and reasoning steps, comparing them to the actual solutions provided. This exercise aimed to encourage self-assessment and identification of errors.

Identification and Critique of Errors

- Correctness Evaluation: The agent assessed the correctness of its solutions, identifying where answers were incorrect or reasoning was flawed.
- Error Analysis: For each problem, the agent detailed the specific errors made, such as miscalculations, misapplications of theorems, or misunderstandings of the problem statements.

- Patterns of Mistakes: The agent recognized recurring issues, such as overlooking key details in problems or insufficient exploration of all possible cases.

- Recommendations for Improvement: Suggestions were made on how to enhance future problem-solving approaches, including double-checking calculations and ensuring thorough understanding of problem statements.

Additional Tasks Assigned to the Agent

Task to Research and Create Strategy Guide

The agent was assigned a complex task: to compile a comprehensive strategy guide and knowledge base for the AI Mathematical Olympiad Progress Prize 2 competition. The task required: • Researching competition details, rules, and guidelines. • Analyzing past competitions and identifying key techniques. • Compiling useful resources and insights from discussion pages.

Instructions Included: • Resource Utilization: Search in-memory resources first, then conduct thorough online research, including reading articles and checking competition discussion pages. • Output Requirements: Produce a detailed, in-depth strategy guide, incorporating insights from research and testing different methods. • Error Handling: Avoid infinite loops, monitor for repeated errors, and ensure environment stability.

Agent’s Performance and Observations

- Initial Attempts: The agent began by attempting to retrieve competition details from memory but encountered limitations due to insufficient in-memory data.

- Online Research Challenges: The agent struggled to effectively search online, sometimes accessing irrelevant websites or failing to retrieve content.

- Limited Depth in Output: The strategy guide produced lacked the requested depth and detail, with sections being superficial and not incorporating insights from articles or discussion pages.

- Failure to Test Hypotheses: The agent did not adequately test different methods or explore hypotheses as instructed.

- Repeated Errors: There were instances of the agent repeating the same approaches without success, indicating a need for better error handling and strategy adjustment.

Analysis of Agent’s Capabilities

Strengths

- Basic Problem-Solving Skills: The agent demonstrated the ability to solve standard mathematical problems, especially when provided with hints.

- Ability to Follow Instructions: When explicitly guided, the agent could include reasoning steps and attempt self-review.

- Identification of Errors: The agent could recognize and articulate errors in its solutions upon review.

Weaknesses

- Independent Problem-Solving Limitations: The agent showed difficulty in solving complex problems without guidance or hints.
- Inadequate Research Skills: The agent struggled to conduct thorough online research and utilize external resources effectively.
- Superficial Reasoning: Reasoning steps were often not detailed enough to fully explain the solutions or demonstrate deep understanding.
- Error Handling Deficiencies: The agent did not effectively adjust strategies when encountering repeated errors, leading to inefficient task execution.
- Task Execution Challenges: In tasks requiring initiative and critical thinking (e.g., creating a strategy guide), the agent did not meet the expected depth and comprehensiveness.

Recommendations and Future Work

1. Enhance Problem-Solving Strategies:
   - Implement training modules focusing on advanced mathematical concepts and problem-solving techniques.
   - Encourage the agent to explore multiple solution paths and consider edge cases.
2. Improve Research Capabilities:
   - Develop the agent's ability to conduct effective online research, including evaluating the relevance and credibility of sources.
   - Incorporate training on summarizing and integrating information from various resources.
3. Strengthen Reasoning and Explanation Skills:
   - Emphasize the importance of detailed, step-by-step explanations in solutions.
   - Provide examples of well-structured reasoning to model after.
4. Implement Advanced Error Handling:
   - Equip the agent with mechanisms to recognize when a strategy is not working and to adapt accordingly.
   - Introduce protocols for adjusting approaches after repeated failures.
5. Enhance Task Execution and Initiative:
   - Encourage proactive identification of subtopics and the appropriate delegation of tasks.
   - Foster critical thinking and initiative in tasks requiring planning and organization.
6. Monitor and Adjust Environment Stability:
   - Ensure the agent can detect and respond to environmental issues, such as infinite loops or system errors, restarting or adjusting as necessary.

Conclusion

The testing and evaluation of the agent revealed a foundation of basic capabilities in mathematical problem-solving and task execution, particularly when guided with hints and explicit instructions. However, significant areas for improvement were identified, including independent problem-solving skills, research proficiency, reasoning depth, error handling, and initiative in complex tasks.

By addressing these weaknesses through targeted training and enhancements, the agent’s performance can be improved to meet the demands of advanced problem-solving and complex task execution. Continued monitoring and iterative development will be essential in refining the agent’s capabilities and achieving the desired level of proficiency.