MCP evaluations/Testing

Objective

The primary goal of MCP testing is to ensure that the MCP server or tool behaves as expected when provided with different types of datasets, including those that should trigger tool calls and those that should not.

Testing Process

1. Dataset Creation

- **Positive Dataset**: Create a dataset containing prompts where tool calling is expected.
- **Negative Dataset**: Create a dataset with prompts that are irrelevant to the MCP tool, ensuring no tool call should be triggered.

2. Manual Testing

- Run the datasets against the MCP tool.
- This can be done using platforms such as **Cursor** or **Claude Desktop**.
- Verify manually whether the tool calls occur as expected based on the dataset type.

3. Automated Testing

- Develop scripts or a framework to run datasets against the MCP tool automatically.
- Capture and log the results for each prompt.
- Compare the results with the **ground truth** or **reference output** to validate correctness.
- Automated Testing Frameworks: Utilize automated testing frameworks to create and execute test cases for both the MCP server/client and the AI model.

4. Evaluation

- Accuracy and Relevance: Assess if the AI model, with the context provided by MCP, can generate more accurate, relevant, and comprehensive responses or perform tasks more effectively compared to a baseline without MCP .

- Efficiency of Context Utilization: Evaluate how efficiently the AI model uses the provided context. Is it able to extract the most relevant information and integrate it seamlessly into its reasoning or actions?
- User Experience: If the AI model interacts with users, evaluate the overall user experience, including the clarity and helpfulness of responses, and the AI's ability to understand and respond to user queries in a contextually appropriate manner.

Validation Criteria

- **Correct Tool Invocation**: Positive dataset prompts should result in accurate tool calls.
- **No False Positives**: Negative dataset prompts should not trigger unnecessary tool calls.
- **Result Accuracy**: Outputs should match the predefined ground truth.
- **Stability**: Repeated executions should yield consistent results.
- A/B Testing: Compare different MCP configurations or AI model versions (with and without MCP) using A/B testing to measure the impact on key metrics.

Conclusion

By following this process, MCP evaluations can be performed both manually and through automation, ensuring correctness, reliability, and robustness of MCP tools and servers.

Scenarios:

1. Connect to MCP server and verify all the tools are available
2. Give a prompt and expected an output and verify correct tool is called
3. Data set creation